STATISTICAL TABLES AND GRAPHS

BY

BRUCE D. MUDGETT

PROFESSOR OF ECONOMICS
UNIVERSITY OF MINNESOTA

HOUGHTON MIFFLIN COMPANY

BOSTON · NEW YORK · CHICAGO · DALLAS
ATLANTA · SAN FRANCISCO

The Riverside Press Cambridge

The Riverside Press
CAMBRIDGE · MASSACHUSETTS
PRINTED IN THE U.S.A.

PREFACE

THIS book has grown out of ten years' experience in teaching elementary statistical methods to students preparing to enter the School of Business Administration at the University of Minnesota. The course has been designed, not for the students who intend later to become statisticians, but rather for those who intend to become business men, and emphasis has been placed upon those elementary statistical methods with which business men are likely to come into most intimate contact. With this group, tables of figures and the graphic methods of representing statistical facts are a very important part of the field of statistical methods. For the more refined processes and methods of analysis, business men must depend on specialists; but tables and graphs have become the tools of large numbers of persons to-day who cannot claim to be technically trained statisticians.

In this situation a teaching difficulty arises, for most of the elementary textbooks treat rather inadequately of these two topics and the many books devoted to graphics are so large as to be unsuited to a course that devotes but a few weeks of the term to tables and graphs. As a matter of fact, it is more than a teaching difficulty, for there are very few business men or men in any other walk of life who, when they want to make use of graphic methods, have the time or patience to wade through a book of several hundred pages to find what they want. This is particularly the case if it is possible to present all the essentials of graphics within relatively much smaller space. The present writer has always contended that most books on graphics are too large, and that the essential subject matter can well be presented in much shorter space; and the present book is an attempt to justify this claim.

The attempt to curtail the presentation of technical subject matter, however, may result in a superficial treatment — a situation to be avoided equally with a treatment that is too extended. The discussion in the succeeding pages has been guided always by the conviction that both table structure and statistical graphics represent unified bodies of technique, the several parts of which are logically related the one to the other; and all of them directed as means to a given end, to present relationships among statistical facts. It may be well to note that the limitation of the title has been adhered to consistently — table structure and graphic methods have been discussed only in so far as they have to do with the presentation of *statistical* data. It is hoped that the logical connections of the subject matter and the logical relationships between technical methods and data to be represented have been kept more in the foreground in this book than in its more voluminous predecessors.

So far as the title-page is concerned the book is the product of one author, and the statement holds with regard to the actual writing. But there has been a contribution of many minds to the logic of the following pages. Needless to say, other writers on graphics have contributed heavily to many of the formulations. But my debt has been heaviest to those who, during the last ten years, have assisted in teaching the course in which the book has come into being, and who in conference and in classroom experience have put to the test practically every device of table structure or of graphic method here discussed. In addition I have had advice and criticism from three members of the staff who have read the manuscript, from Mrs. Nina L. Youngs, Mr. Richard Kozelka, and Mr. H. G. Fraine.

BRUCE D. MUDGETT

ANALYTICAL TABLE OF CONTENTS

PART I. STATISTICAL TABLES

PART II. GRAPHIC STATISTICS

STATISTICAL TABLES AND GRAPHS

PART I

STATISTICAL TABLES

STATISTICAL TABLES AND GRAPHS

I

THE CLASSIFICATION OF STATISTICAL DATA

The nature of statistical data. The social sciences present an especially difficult problem to the scientific investigator because of the complex character of most of the data with which he must work. The experimental method, which is the characteristic method of the physical sciences, can be used only to a limited extent in the social sciences, because of the impossibility of controlling the circumstances under which experiments are made. When the physicist wishes, for instance, to demonstrate the action of gravitation, he finds it possible to set up an experiment under which a falling body is influenced to no important extent by any other than the attractive force of gravitation. But the economist who wishes to ascertain the influence of increased production upon the price of a commodity entertains no prospect of being able to set up his experiment in a vacuum from which all important influences other than that of production increase are excluded. He must go to the market-place where the good is bought and sold and must observe what takes place there. The influences at work in the market-place, however, are not of his choosing, and they may include many things other than production changes. The price changes which he observes result from the combination of all these influences, and his task is to devise a method of analysis that will bring about a cancelling of the effects of the *other* influences and leave a residuum that can reasonably be attributed to the particular force that is being studied.

It is to deal with situations such as the one just described that the methods of statistics have been developed, methods which are applicable to mass data and which make possible some progress in separating observed complex results and in relating the separate parts to specific causal factors. And so one author [1] has defined statistical methods as "methods specially adapted to the elucidation of quantitative data affected by a multiplicity of causes." And statistical data, then, represent these "quantitative data affected by a multiplicity of causes."

To consider this definition in greater detail, the price and production illustration just used may be considered further. The price of wheat may be observed, for instance, on a given day in a given market, and there may be adequate reasons for the assertion that this price is related to, or affected by, the wheat crop of the current season. At the same time it cannot be asserted that there is a direct relationship between the observed price and the crop similar to that which exists between the speed of a falling body and the attraction of the earth. The wheat price is simultaneously affected by many things other than the size of the crop — by the supply of, and demand for, other products that may become possible substitutes for wheat; by the supply of money in relation to the demand for it (i.e., the price-level); by temporary relationships existing between traders in the immediate market; and by other forces unnamed or unnamable.

Of the various phenomena that seem to offer logical explanations of price fluctuations, some are more important in one situation or at one time than others. For instance, on a given day the price of wheat may not be as greatly affected by changing world crop conditions as by forces that can work out their effects during the day, such as the temporary credit situation of the traders in the market. On the other hand, if one considers the level of wheat prices for a season, these daily credit condi-

[1] Yule, *Introduction to the Theory of Statistics*, 6th edition, 5.

tions are of slight importance, while the size of the crop becomes a major factor. Now consider the fact that the price observations which one studies are the prices at which actual exchanges take place from moment to moment in the market. Given a record of these prices for a year, the problem is to discover from them the effect of influences operating within any given day and of influences which persist throughout the season. Obviously, a first step will be to separate, or to classify, the observations in such a way that each day's record will stand by itself and then one day's record may be studied or compared with another. Any given day will show many transactions, if there is a large and active market, but the average price at which these transactions have occurred will ordinarily show the effect of influences that have persisted throughout the day; and a comparison of such averages for two consecutive days will show whether important day-to-day forces have been more active one day than another. It may happen, for instance, in the speculative wheat market that there is a heavy "long" interest in the market on margined accounts; that is, many people may have bought wheat in the hope of a rise in price and have borrowed from banks or brokers a large proportion of the purchase price. Other traders sensing this situation will sell the commodity short, thereby attempting to depress the price, and if successful may bring about forced sales of the margined accounts. The result will be a lower average price for the day than had existed previously when the market was dominated by bullish influences. The taking of statistical averages of the actual prices does not, it must be emphasized, prove the existence of these day-to-day price-making forces, but if their existence can be established on adequate logical grounds, the statistical averages will measure their effect upon prices.

The longer-time, or seasonal, forces may likewise be studied in terms of price averages and offer further clarification of the nature of statistical procedures. Either of two methods may

be followed over the season in order to trace the relationship of price to size of crop: (1) the entire set of actual prices for the year may be combined and an average taken and this average set against production for the year; and a comparison with similar price and production figures of previous years will indicate whether variations in production are accompanied by similar or opposite variations in price. The question now naturally arises how the variations in these average prices can be related causally to production variations, since the averages are obtained from actual observations that are affected by many other causal factors including the short-time factors considered above. The answer is that the averages are not affected by the other forces in the same way as are the actual observations. To illustrate in the case of the day-to-day forces, on one day the price average will fall compared with the previous day because the market is in control of the "shorts," but on some other day the price average will rise due to the greater activity of the "longs"; and in the course of a season it is highly probable that these short-time rises and falls will about cancel each other and therefore leave no, or very little, effect on the seasonal average. The problem of isolating the effect of a seasonal force, in this case, finds its solution in "averaging out" or cancelling the effects of non-seasonal factors.

(2) Another method of isolating and studying the influence of seasonal causes upon wheat prices consists in grouping or classifying the observed prices into periods longer than a day but sufficiently short to obtain several groups for the season, say by months, and taking averages for each month. The monthly averages would tend to effect the cancellation of daily influences, and the movement of the twelve monthly averages would tend to show the effect of influences that developed as the season advanced.

Necessity for classifying statistical observations. These two influences, seasonal and daily, do not begin to exhaust the list of

factors affecting wheat prices, but the illustration serves as an example of the character of statistical data, as observed phenomena affected by a *multiplicity of causes*. The methods of statistics do not determine causes, but when the causes have been discovered or enumerated by any logical process, their influence upon the observed data may frequently be measured or evaluated by appropriate devices of statistical analysis. In the illustration used, the beginning of this analysis lay in determining certain time categories within which the causative forces operated and in classifying the observations into these categories, thus making possible a study of the observations within each category and a comparison of one category with another. This process of classification is always a second step in statistical analysis to follow immediately upon defining the field of study and gathering the observations for study within the field. The routine process of classifying data is called tabulation.

Kinds of classification. The illustration has involved classification of the observations into time classes or groups, but it is probable that before all classifications of the price data were completed into which the study of various causal relationships of price might lead one, it would be necessary to consider other bases of classification than time. Assuming for the purpose of the illustration that any information desired with reference to actual sales of wheat were obtainable, the purposes of the price analyses might be furthered by knowing whether sales or purchases were made by speculators as such (i.e., professional operators or the so-called "public") or by dealers as hedging transactions. Or it might be desirable to know whether sales were for cash or on margin; or again, the distinction might refer to different sections of the market area in which sales originate, whether in the east, west, north, or south. Each kind of classification that may be suggested may be included under one of four descriptive titles: (1) chronological, or time; (2) geographi-

cal, or areal, or spatial; (3) quantitative, or magnitudinal; and (4) qualitative. Most of these titles carry their own definition. To illustrate, time classification may be by day or week or month or year or any other recognized subdivision of passing time; spatial classifications may be by political divisions, such as State or county, or by any other division of area suited to the purpose in hand — e.g., division of a city by its department of health into sanitary districts or division of a sales territory into districts to fit the needs of a business firm; quantitative subdivision of observations is illustrated by classification of cities or of sales districts by size, human beings by weights, heights, ages, etc. The qualitative classification is one that requires specific definition of each of its categories or classes. In some cases these classes are easily differentiated by all persons who are likely to use them and the setting up of such a classification involves no especial difficulty. Cases in point are divisions of the population into sexes, classification of sales by sales persons or by departments and of wheat produced into spring and winter wheat. But in many cases the problem of defining or delimiting the various categories of the qualitative classification is a difficult and troublesome one. Consider, for instance, the problem of differentiating the various color classes of the population in such a way that a census enumerator will make no mistake as to whom he classifies as white or black or mulatto; or the more difficult problem of classifying the population by nationality — whether a Russian or a Rumanian or a Canadian always represents a distinct ethnic group; or the still more complicated task of differentiating population groups by occupation. The Census Bureau has been working for more than a quarter-century on this problem, and there were included in the Fourteenth Census index of occupations over twenty thousand occupational designations and five hundred and seventy-two distinct occupations or occupational groups.[1]

[1] *Fourteenth Census*, IV, 12.

Cross-classification (or cross-tabulation). The purpose of a given analysis of statistical observations may be served by a single classification, but it is more frequently the case that more than one will be needed. Thus, in studying the growth of a business over several years for the purpose of explaining why at some time the growth has been rapid, at other times slow, it is possible that the manager may obtain all the information he needs by arraying the total sales by months or years; or he may find valuable additional information by so arraying the figures for each sales district, and it is possible that the latter arrangement will show that important information had been masked by the classification of total sales alone. The accompanying sketch illustrates the character of the two procedures. The

Year and Month	Total Sales $
1926	....
Jan.	....
Feb.	....
Mch.	....
etc.	etc.

Year and Month	Amount of Sales in Dollars			
	Total	First District	Second District	Third District
1926				
Jan.				
Feb.				
Mch.				
etc.				

first classification gives total sales alone for each month or year, while the second gives, in addition, the monthly and yearly sales in each sales district. This arrangement is called a cross-classification or double tabulation. The vertical column at the left, called the stub, contains in this instance all the time categories for which data are to be included; the classification across the top of the larger sketch, called the caption, includes a complete list of geographical categories. Any one row of the table therefore gives a geographical classification for one time

category; any one column, a time classification for one geographical category. It is this feature that is referred to in calling the arrangement in the second sketch a cross-classification or a double tabulation. The stub and the caption contain separate and independent classifications. If, for instance, sales data were available by months for several years and were given in a table in which the months, January to December, were listed at the left, while the years, say 1920 to 1928 were listed at the top of the table, this would be a convenient method of listing monthly observations for nine years, but would not be a cross-classification, since both caption and stub refer to the same classification and the purpose of the arrangement is to save space.

To return the discussion for a moment to the problem of causal relationships, the causes operating upon sales development may all require time to work out their effects, but there may be a different set of causal factors in operation in one sales district from those in another. The classification of total sales by months, therefore, fails to disclose these differences between districts. Stated in the technical phraseology of statistics the data of the various sales districts are not homogeneous (i.e., similar with regard to the problem in hand) and cross-tabulation serves the purpose of reducing the data to more nearly homogeneous groups, which may then be subjected to further analysis by averaging or comparison. Homogeneity, it should be remarked, is not an *absolute* quality of statistical materials, but is always *relative* to the problem in hand. For some purposes a given set of observations may be homogeneous and for other purposes heterogeneous. Cross-tabulation, then, aims to arrange the observations into groups which are similar; and the nature of this similarity needs to be examined further. To say that a set of observations are homogeneous is not to say that they are not subject to variations from a large number of causal forces, but rather to say that the majority of these forces affect

the observations at one time in one direction and at another time in the opposite direction, or tend to make the values of the observations larger in some instances, smaller in others, so that for fairly large groups of cases the effects of these many small forces tend to cancel and to leave the average unaffected. The average, then, is determined essentially by the predominant causal factor, or factors, acting upon the observations. And it is only when the observations are thus homogeneous that the average becomes a stable value and permits one to draw safe conclusions from it or by comparison of it with other averages. Conclusions based upon averages of heterogeneous data are dangerous and untrustworthy.

Cross-tabulation is, then, a means of reducing statistical observations to usable proportions. The possibilities of the method are not limited to double tabulation. There may be triple, quadruple, or even quintuple tabulation; though the attempt to work out a complete tabulation scheme for more than three or four variables soon reaches a stage of complexity that becomes almost unworkable. It is not that complete tabulation of five or more variables is impossible, now that we have mechanical equipment to do the routine work, but rather that it is difficult to put the data of a five-variable tabulation into a statistical table and to see its significance when it is so tabulated. If for the dollar sales classified above by months and by sales districts we also had data on each sale indicating the commodity sold and the terms of sale, whether cash or credit, this would give four variable factors stated with reference to each sale — time, district, commodity, and terms. Consider a tabulation of all these data for the year's business. Recalling that the double tabulation required a complete classification of the monthly data *for each sales district* — i.e., all time subdivisions for each area subdivision — a triple tabulation similarly requires for each class or group of the third variable a complete double tabulation of the other two. A triple tabulation, there-

TABLE 1.

CLASSIFICATION OF MONTHLY SALES FOR 1926

BY DISTRICTS AND CLASS OF GOODS

(in dollars)

DISTRICTS AND CLASS OF GOODS	MONTHS											
	Jan.	Febr.	Mch.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
TOTAL SALES												
China												
Hardware												
Hosiery												
Men's Furnishings												
Millinery												
Overcoats												
Shoes												
Sporting Goods												
Suits — Men's												
Toys												
FIRST DISTRICT												
China												
Hardware												
Hosiery												
Men's Furnishings												
Millinery												
Overcoats												
Shoes												
Sporting Goods												
Suits — Men's												
Toys												
SECOND DISTRICT												
China												
Hardware												
Hosiery												
Men's Furnishings												
Millinery												
Overcoats												
Shoes												
Sporting Goods												
Suits — Men's												
Toys												
THIRD DISTRICT												
China												
Hardware												
Hosiery												
Men's Furnishings												
Millinery												
Overcoats												
Shoes												
Sporting Goods												
Suits — Men's												
Toys												

fore, multiplies the work of a double tabulation by a factor equal to the number of classes in the third classification. A year's sales by months classified for each of three sales districts, as above, page 9, requires a table with thirty-six cells or boxes for the statistical data, not including totals. If now a threefold tabulation is made by indicating also kind of goods sold, and if the latter classification shows ten kinds of goods, the threefold tabulation will require a table containing ten times thirty-six or three hundred and sixty cells. Table 1 is prepared for the data of such a threefold tabulation and, with the totals included, actually provides for five hundred and twenty-eight separate entries ($12 \times 11 \times 4$). If now a quadruple tabulation were contemplated to include also the classification of these sales by terms of sale, cash or credit, the size of the above table would be doubled and its unwieldy character would become apparent. So far as the logic of tabulation procedure is concerned, it would probably be more satisfactory, where the fourfold classification is desired, to present the results in two threefold classification tables, one for cash sales and one for credit sales, or possibly in three triple-classification tables, one for each sales district showing cross-classifications of data by months, kinds of goods and terms of sale. It is evident that a great number of combinations might be made and the selection would depend on the purposes the various tabulations were intended to serve. Assume that the sales manager's main interest lies in the growth of sales during the year; that he wishes to have these figures not only for the business as a whole but for different kinds of goods, and that he wishes to compare growth through the year by districts and for cash and credit sales. The accompanying sketches are suggestive of the tabulations that would satisfy his requirements. Development of sales through the year being a primary interest, the classification by months is in each instance placed in the stub. The first sketch permits a comparison of totals by months and the same for each of six departments or kinds

(1)

MONTHLY SALES 1928, BY KINDS OF GOODS

MONTH	TOTAL SALES	CHINA	HARDWARE	HOSIERY	MEN'S FURNISHINGS	MILLINERY	OVERCOATS

(2)

MONTHLY SALES 1928, BY DISTRICTS AND TERMS OF SALE

MONTH	TOTAL SALES		DISTRICT 1		DISTRICT 2		DISTRICT 3	
	Cash	Credit	Cash	Credit	Cash	Credit	Cash	Credit

(3)

MONTHLY SALES 1928, BY DISTRICTS AND TERMS OF SALE

MONTH	TOTAL SALES			CASH SALES			CREDIT SALES		
	District 1	District 2	District 3	District 1	District 2	District 3	District 1	District 2	District 3

of goods. The remaining sketches offer two arrangements of a triple tabulation, time, district, and terms, the first of which would be selected if the greater interest attached to the comparison of cash and credit sales as opposed to comparison of districts; the second being preferable if the interest in district and terms of sale were reversed.

In any instance, a final selection from the data may include one or more single, double, or triple tabulations. The point to be emphasized is that complete tabulation is never, in itself, a desideratum. Tabulation is intended to serve the purpose of the analysis, and no details need be given that are not directed toward that end.

II

THE PROCESS OF TABULATION

THE routine process of distributing the observations according to the various classifications and cross-classifications decided upon may be done by hand, but is more conveniently and more economically performed by mechanical tabulators whenever the tabulation includes a large number of cases with a considerable number of facts for each case. With a hundred or possibly a few hundred observations to tabulate and a relatively small number of classifications to make, the work can probably be done more cheaply by hand, but if the cases run to a thousand or more, or if the number of classifications is large, mechanical methods are likely to be cheaper. The point at which the one method will surpass the other in economy will vary for different investigations, but can be estimated with a fair degree of accuracy upon the basis of brief experience with tabulations.

First steps: preparation of the classifications and of the statistical tables. Two steps regularly precede any work at actual tabulation. In the first place the subdivisions of the various classifications must be decided upon. Frequently this work needs to be done before the observations have been assembled in order that the information collected will furnish an adequate basis for distribution of the items in the classification. The development of a new and improved classification may necessitate a complete rephrasing of the questions asked in the original schedule. For example, at the thirteenth decennial census an entirely new classification of occupations was introduced, and whereas previous population census schedules contained but one column for the designation of occupation, the schedule for 1910 had two columns, one for "trade or occupation," the other

for "industry," since the new classification involved a trade or occupational grouping within an industry framework.[1] The "cause of death" question on the standard death certificate has also been changed as a result of the development of a new classification of causes of death. The division of a city into sales territories by wards might be much more unsatisfactory for the purpose of developing sales (discovering new sources of demand) than a division based roughly upon income strata or upon broad occupational groupings. And likewise a traditional classification of departments based largely upon the historical development of the business might not be as valuable as one which sought a grouping of goods which classed together those which appealed to similar groups of purchasers.

Though the problem of determining subdivisions of classifications is generally simpler for geographic, quantitative, and time groupings than for those of the qualitative type,[2] there are left difficulties aplenty in selecting the given classification of any type that will lead most directly to the discovery of facts and relationships of greatest value in the analysis. The important consideration for the technical process of tabulation is that all these questions of classification must be settled before tabulation begins, for they are a significant link in the chain of factual analysis that is to lead to sound conclusions; they represent the logic of the statistical method, while the tabulation itself is a clerical task.

It is similarly important, before tabulation begins, to have decided upon all the classifications and cross-classifications that are to be made in order that the details of tabulation shall not be carried to unnecessary lengths; and in furthering this end it is necessary that all table-forms into which the classified data are to be placed shall be drawn up in advance.

Hand tabulation. The observed data are sometimes gathered on small cards suitable for repeated handling, and when this is

[1] See *Fourteenth Census*, IV, 10–22; 27–29. [2] See page 8.

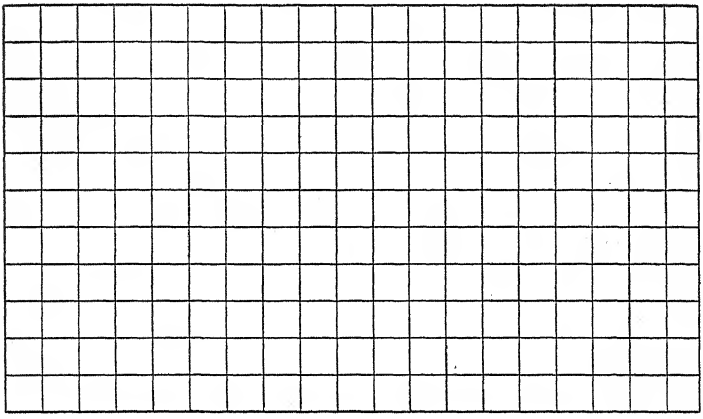
card for schedule.

the case the classifications may conveniently be made by sorting these cards into the appropriate groupings and repeating the process as often as necessary for different classifications. It may be desirable also to make up a rough work-sheet showing the classifications and cross-classifications needed, so that the counted frequencies that result from distributing the cards into each class can be recorded and checked before being placed finally in the tables. The cards will ordinarily need to be edited before the process of distributing into classes begins in order to indicate with certainty in each instance into which class the recorded data fall.[1]

Where several cross-classifications are to be made, it is best to distribute the cards first in that classification which occurs most frequently; then to take each group in this classification and distribute the cards in it according to the groupings of a second classification. If, in a sales analysis, for instance, each table that is to be prepared contains the classification of sales by departments — i.e., kind of goods — it would be desirable to distribute the cards first by departments. Having done this, it would then be important to look carefully over the cards in each group to check against errors in misplacing, then to count and record on the work-sheet the number of cards in each group and compare the totals with the known number of cards in the investigation. This matter of checking the accuracy of the clerical work needs to be emphasized, for it is very easy to make mistakes in it and a careless person can make enough of them to invalidate the conclusions that may be drawn. If now it is desired to cross-classify sales by departments with sales by districts, the procedure will be to take the groups of cards as now distributed by departments and for each department separately to distribute them by sales districts. In case the cards are small and a fairly large table is available to work on,

[1] This is aside from, and in addition to, the editing which is done in order to check accuracy in recording the original observations.

it may be worth while to have the top of the table marked with a series of lines to form rectangular spaces into which the cards can be distributed after the two classifications have been inserted, one at the top and one at the side of the table, as illustrated herewith:



The classifications may be written into the table with chalk or crayon and then erased when the distribution is completed. Where a threefold classification is to be made, the cards in each cell or rectangle of this table may then be distributed into the subdivisions of the third classification.

More frequently than not the original observations are on sheets of a character such that they cannot be distributed as described for the cards. For many analyses made by business firms the data are the facts given on sales-slips, order-forms, invoices, and the like. These are so large and unwieldy or the paper is so thin that it is impossible to distribute them into groups and count them. The desired data, therefore, must be transferred from these records to work-sheets, or tally-sheets one record at a time. In the cross-tabulation used in the illus-

tration above, it would be desirable to make up a tally-sheet showing the same arrangement as in the lined table, or this table could be used if it is not too unwieldy and if tally-lines could be made distinctly upon it. Then the necessary facts from each record could be tallied by any simple process, such as

finally counted for each subdivision, and recorded in the tables. Extreme care is necessary in thus transferring the record, for there is no convenient means of checking the accuracy of the work unless the same classification is tallied a second time and the two results agree.

Mechanical tabulation. In order to distribute the data into the various classifications by mechanical methods, it is necessary that a set of code numbers be arranged for each classification so that each class or subdivision of the classification shall be represented by a separate number. Thus, if a firm have not more than ten sales districts, they may be represented by the numbers 0 to 9 inclusive; a classification including one hundred items would require the numbers 0 to 99. If a given classification were required to show dollar values of sales, the code would require a series of numbers large enough to present the largest item of dollar sales. On pages 21 and 22 are given parts of two codes used by the United States Shipping Board in 1919 in the tabulations of its operating division. A single page is shown from the port code and one from the classification of accounts—sufficient in each instance to show how code numbers are made to represent the items of a detailed classification. Having arranged the code numbers for each classification, the coded data are then transcribed from the original schedule or blank upon which they have been collected to cards suitable for use in the sorting and tabulating machines. Illustrations of these cards are given on pages 23, 24. Two of them are printed in blank, the third divided into fields for a sales analysis. One type of card is composed of forty-five columns, each containing the numbers

FIRST REVISED ISSUE
May 1, 1919

U. S. Shipping Board — Division of Operations
Tabulating Section — Office of the Comptroller

DETAILED CLASSIFICATION OF ACCOUNTS

Code Number	Description
	OPERATING EXPENSES
	DECK DEPARTMENT
100	WAGES
101	Officers
102	Petty Officers
103	Seamen
104	Cadets
105	Wireless Operators
106	Messmen and Mess Boys
107	Carpenters and Electricians
108	
109	
111	Bonus
112	Overtime
113	Board Allowance
114	
115	
120	SUPPLIES
121	Nautical Instruments, Charts
122	
123	Ropes — Wire and Manila
124	Canvas
125	Hardware and Tools
126	Paints and Paint Brushes
127	Cleaning Articles
128	Flags, including Signal Flags
129	Containers
130	
131	Lamps — Electrical and Oil
132	Oils for All Deck Uses
133	Marlin Spikes, Shackles, Tackle Blocks and Falls
134	
135	
138	Other Supplies
140	EXPENSES
141	Rating Chronometer, Adjusting Compass
142	
143	
	ENGINE DEPARTMENT
200	WAGES
201	Officers
202	
203	Oilers and Wipers
204	Cadets
205	Firemen, Water Tenders, Pumpmen
206	Messmen and Mess Boys
207	Coal Passers
209	
210	
211	Bonus
212	Overtime
213	Board Allowance
214	
215	
229	Containers
230	
231	Electrical and Oil Fixtures
232	Waste and Curled Hair
233	
234	Pipe Fittings — Valves, Couplings
235	
236	Grate Bars
237	Fire Bricks, Cement and Clay
238	Other Supplies
239	
240	EXPENSES
241	Scaling of Boilers
242	
243	Removal of Ashes
244	
245	
250	FUEL
251	Coal
252	Cost of Coaling
253	Fuel Oil
254	Fueling
255	Cost of Steam
256	
257	
	STEWARD'S DEPARTMENT
300	WAGES
301	Steward and Storekeeper
302	
303	Cooks
304	
305	Cabinboys
306	Messmen and Mess Boys
307	
308	
311	Bonus
312	Overtime
313	Board Allowance
314	
315	
320	SUPPLIES
321	Medicines, Medical Attention
322	Water, Drinking
323	Fuel (Cooking) and Matches
324	Linens
325	Tableware, Glass and China
326	Tableware, Silver
327	Cleaning Articles
328	Bedding
329	Containers
330	Bar Supplies
331	Kitchen Utensils
332	Ice
333	Wearing Apparel

PORT CODE
100 ATLANTIC NORTH AMERICAN REGION

110 CANADIAN DISTRICT (including the province of Quebec and Newfoundland)

- 110 All other ports
- 111 Montreal, Quebec
- 112 Quebec, Quebec
- 113 Three Rivers, Quebec
- 114 Tadousac, Quebec
- 115
- 116 All other St. Lawrence River ports
- 117 St. Johns, Newfoundland
- 118 Grand Bay (Port-Aux-Basques), Newfoundland
- 119

120 GREAT LAKES DISTRICT

- 120 All other ports
- 121 Buffalo, N. Y.
- 122 Cleveland, Ohio
- 123 Detroit, Mich.
- 124 Chicago, Ill.
- 125 Duluth, Minn.
- 126 Milwaukee, Wis.
- 127 Toledo, Ohio
- 128 Toronto, Ontario, Canada
- 129

130 MARITIME PROVINCES DISTRICT (including New Brunswick, Nova Scotia, and Prince Edward Island, Canada)

- 130 All other ports
- 131 Halifax, N. S.
- 132 Sydney, N. S.
- 133 Louisburg, N. S.
- 134 Pictou, N. S.
- 135 Yarmouth, N. S.
- 136 S. John, New Brunswick
- 137 Chatham, New Brunswick
- 138 Moncton, New Brunswick
- 139 Charlottetown, P. E. Is.

140 NORTHERN NEW ENGLAND DISTRICT (including Maine and New Hampshire)

- 140 All other ports
- 141 Portland, Maine
- 142 Portsmouth, N. H.
- 143 Bangor, Maine
- 144 Bath, Maine
- 145 Belfast (Searsport), Maine
- 146 Rockland, Maine
- 147

160 NEW YORK DISTRICT (including the port of New York and associated ports in New Jersey)

- 160 New York Harbor (not classified)
- 161 Manhattan, N. Y.
- 162 Staten Island, N. Y.
- 163 Brooklyn, N. Y.
- 164 Hoboken, N. J.
- 165
- 166

170 MIDDLE ATLANTIC DISTRICT (including ports in New Jersey not contiguous to New York and the ports of Pennsylvania, Delaware, Maryland and Virginia)

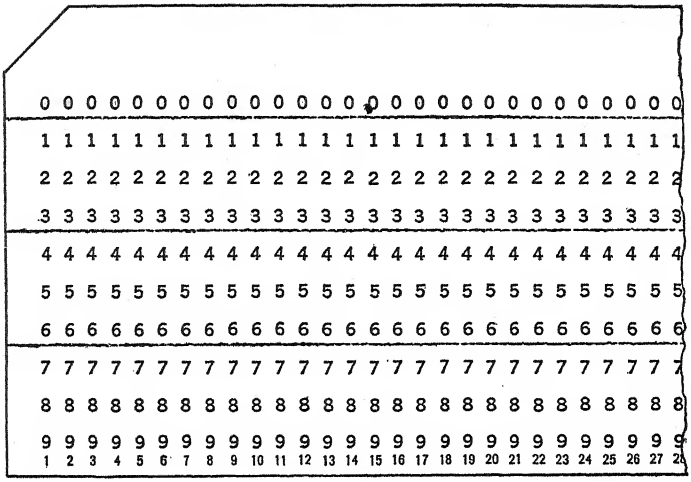
- 170 All other ports
- 171 Philadelphia, Pa. (Camden and Chester)
- 172 Norfolk, Va. (includes Portsmouth, Berkley and Newport News, Va.)
- 173 Baltimore, Md.
- 174 Wilmington, Del.
- 175
- 176
- 177

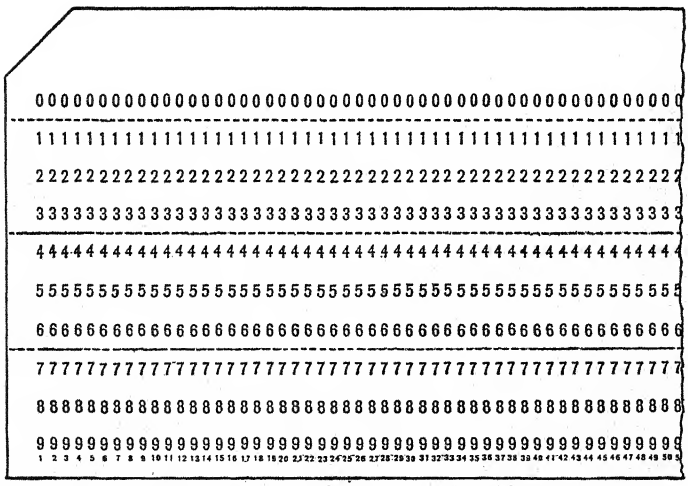
180 SOUTH ATLANTIC DISTRICT (including North Carolina, South Carolina, Georgia and eastern Florida and Bermuda Islands)

- 180 All other ports
- 181
- 182 Wilmington, N. C.
- 183 Georgetown, S. C.
- 184 Charleston, S. C.
- 185 Savannah, Ga.
- 186 Brunswick, Ga.
- 187 Jacksonville, Fla.
- 188 Hamilton, Bermuda Is.
- 189 All other Bermuda Is. ports

190 GULF DISTRICT (including western Florida, Alabama, Mississippi, Louisiana and Texas)

- 190 All other ports
- 191 New Orleans (Port Chalmette), La.
- 192 Galveston (Pt. Bolivar, Freeport), Texas
- 193 Sabine (Beaumont, Port Arthur), Texas
- 194 Mobile, Ala.
- 195 Pensacola, Fla.
- 196 Tampa, Fla.
- 197 Boca Grande, Fla.
- 198 Key West, Fla.
- 199 Gulfport, Miss.





BLANK CARDS FOR USE IN MECHANICAL TABULATION

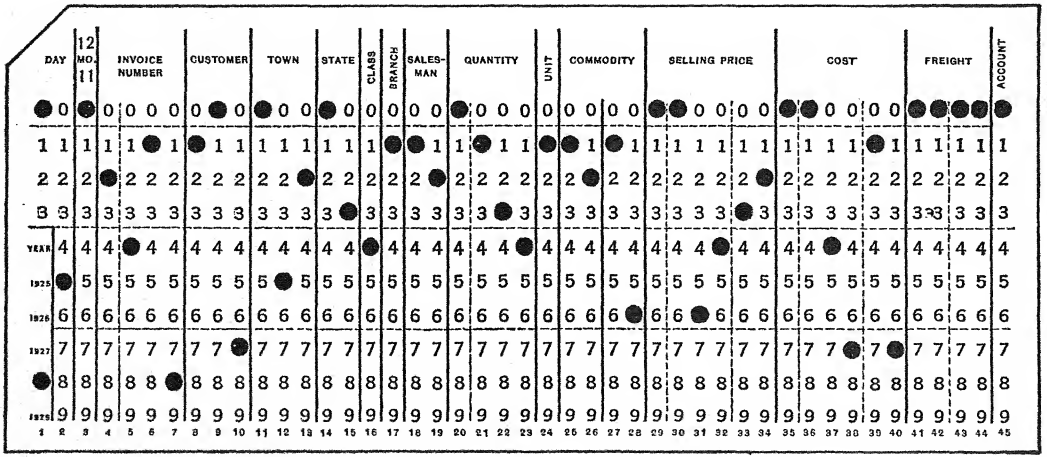
Courtesy The Tabulating Machine Company

| | | |

0 to 9 arranged consecutively from top to bottom of the column, the other of eighty such columns. A single card will contain all the variable facts on one unit or observation. To continue with the sales-analysis illustration, if the data gathered cover kind of goods, sales district, date of sale, and dollar value of sale, each card will contain four coded items, one for each of the above classifications. In many cases a given order-blank or invoice from which the firm obtains these data will contain several items of merchandise which will fall into different classes in the "kind of goods" classification. In this case it is necessary to have a separate card for each "kind of goods" in the given order. In other words, each coded card will contain *one* item of data for each and every classification involved.

The cards in question are prepared for use in a given analysis by being divided into *fields*, one for each classification. Suppose in the illustration above that the kind of goods classification includes thirty items, sales districts ten, date of sale twelve items to designate the month and ten to designate the year; and suppose that value of items sold never ran in any instance above five thousand dollars and values are to be shown in dollars and cents. The card then will be divided into four fields. The first, kind of goods, will require two columns on the card, since two columns are necessary to cover thirty items; the second field, sales districts, needs only one column; the third, two columns for months and one for years. In some cases it is possible to put twelve items of a classification in one column,[1] and where this is done but one column is necessary to designate the month of sale. The fourth field, value of goods sold (or selling price), will require six columns. Thus the four fields involved will use approximately only one quarter of the card. The designations of the various fields on the cards are printed at the top of the appropriate columns, and in a sales analysis such as described the cards so printed can be used for a ten-year

[1] See illustration on page 26, where this is done.



A Punched Sales Card

Courtesy The Tabulating Machine Company

period. The sales analysis card given in the illustration on page 26 shows the coded facts for one sale punched on the card.

Having set the code numbers for each classification and prepared the cards, the task of performing the tabulations involves several steps. First it is necessary to prepare the original record for transferring the information to the card. In some cases this will involve writing the code numbers on the record; in other cases, and particularly where the classification is short, the items of the classification will need to be before the operator who makes the transfer. The transfer is accomplished by placing the card in a machine, called the punch-card machine, and punching out the code numbers in the proper columns of the various fields to indicate the information on the original record. It is important that this transfer of the original data to the punched card shall be carefully checked. This is sometimes done by having a duplicate card punched by a different operator and then comparing the two cards to see that the punched holes agree.[1] The cards having been punched and checked are ready for insertion into the sorting machine, the purpose of which is to distribute them by classes in each classification. The mechanics of this operation is performed by passing the cards through the machine, in the course of which passage a steel brush passes across the face of the card and the machine is so set that when the brush passes over a punched hole in a given card it makes an electrical contact which routes the card into a given compartment. There being ten different positions in each column, any one of which may be punched, there will be ten compartments to receive the cards and ten different electrical contacts that the brush may make in passing over the face of the card. To distribute the cards in a classification that involves more than ten subdivisions, therefore — for instance, the thirty in one in-

[1] The various ways of performing these and other checks necessary in mechanical tabulation need not be explained here in detail, for they involve mostly routine operations and are always carefully explained by the companies selling the tabulating equipment.

stance above — will necessitate running the cards through the sorting machine a second time. By continuing this process as often as necessary, any sortings or classifications of the cards that are desired may be made.

When the cards have been sorted into the divisions that are desired, the next step is to count the cards in each division and to record totals. For the purpose of counting and totaling, the cards are passed through a tabulating machine. The mechanical principle involved is the same as for the sorting machine. A brush which makes various electrical contacts passes over each column containing information to be counted or itemized and by means of this contact the item is recorded on a counter-dial. There is this difference between the tabulator and the sorter, that, whereas the sorting can be done only one column at a time, the recording or counting may involve a large number of columns in one run of the cards through the tabulator.

The great advantages of mechanical tabulation lie in the large number of classifications that can be made in very short time and with small expense, once the data have been punched on the cards. When the classifications have been coded and the original records have been prepared for the punch-machine operator, a skilled operator can punch from two thousand to thirty-five hundred cards a day; and having been punched, the cards can be passed through the sorter at the rate of from two hundred and fifty to four hundred cards per minute, and through the tabulator at the rate of from ninety to one hundred and fifty per minute. It is easy to see from this statement that, when the number of records or observations runs into the thousands and the number of classifications is large, the economy of mechanical tabulation may far surpass that of hand methods. Mechanical tabulation has meant, in the United States Census Bureau, the possibility of vast increases in the scope of the tabulations along with far greater speed in supplying to the public the results of periodic censuses, and all with a greatly decreased chance of clerical inaccuracies.

III.

PRESENTATION OF DATA IN TABLES
CONSTRUCTION OF TABLES

The problem one of presentation of quantitative relationships. The classifications required to obtain the desired information from the observations having been decided upon, and the data having been tabulated in these classifications and in the cross-classifications that are deemed important, the next step is to construct statistical tables into which the results of the tabulations can be placed. Statistical tables are an important factor in making easily understandable the meaning and significance of the tabulations. Their importance arises through the rather common difficulty most people have in reading lists or tables of figures. Many persons, finding tabular material in books, reports, or documents, will shun the tables entirely and will depend on what the writer has taken from them and incorporated in the text material, whereas conclusions are likely to be safer when based on first-hand study of the entire tabulation than when accepted second-hand from another's interpretation. Hence the problem of the construction of statistical tables is a problem in *presentation*. The table-form adds no new meaning to the figures; it only makes the meaning which is already there easier to find. And this is especially important in elementary statistical analysis, for the tables of figures are probably used far more frequently by the non-specialist than by the specialist. Maybe it is not going too far to say that statistical tables should be so designed that *he who runs may read*, that the busy man, who wants the results of statistical analysis and yet who will not take the time for detailed study, may yet receive a correct understanding of the facts.

In the consideration of statistical tables it must never be forgotten that anterior to the construction of the table is a set of relationships to be shown among quantitative data — a classification of a set of observations, or a cross-classification, or a comparison of the relationships between two classifications; and that the discovery of this set of relationships has been the purpose of the analysis. The statistical table is correct just in so far as it has aided in making clear this situation. This central fact must be borne in mind at all times, in the construction of the table and in judging it when finished. The table is not an end in itself, but a means to an end, and that end is the display of the relationships in question, this being at least a part-solution of the general problem involved.

Differentiating general-purpose from special-purpose tables. These introductory statements are made with especial reference to *analysis* tables, or special-purpose tables, as they are sometimes called. They are generally distinguished from repository, or general-purpose, tables, and the distinction is important when considering the uses to which the two types are put, though not in all respects with reference to mode of construction of the tables. The descriptive terms used indicate the difference between the two types, the general-purpose table being designed as a repository of the tabulations in full detail; whereas the analysis table is intended, as the name suggests, to present the results of analysis, to give not necessarily or always full detail, but summaries or conclusions and significant relationships. The distinction is excellently illustrated by the various tabulations of the Federal Census Bureau. The materials of the periodical Federal censuses are gathered to supply the needs of many groups with widely varying interests. The population statistics are used, for instance, as a basis for determining representation in Congress; a great number of business firms or agencies now use them in studying the development of markets; and scientists in various fields, especially economists and soci-

ologists, use them for a variety of purposes. With this wide divergence in interest among all the users of population statistics, it is not possible, within the limits of its restricted budget, for the Census Bureau to organize its population counts in such a way as to meet all the specialized needs of each interested group. The population census must be taken with a *general* purpose in mind — to satisfy the greatest number of uses of the greatest number of users. There are certain classifications of the population so generally demanded that they are included in the population statistics always and without question; others, which are demanded by a few only, may be questioned, and, if the cost is high or if tabulations of more general interest have used up a large share of the budget, they may have to be omitted. These tabulations of general interest are therefore given in great detail in the Census volumes. A good illustration is given in volume 2 of the 1920 Census, the general *Report on Population,* table 13, cross-classifying age distribution with sex, color, and nativity classifications and showing these facts for the entire population, for broad area divisions, such as New England States, Middle Atlantic States, etc., and finally for separate States. This one table, number 13, covers one hundred and sixteen pages of the *Report,* pages 170–285 inclusive. An important characteristic of such a table must necessarily be ease of access. The material must be arranged in such a way that, despite the great amount of detail involved, the user can turn readily to that part of the table that will give him the specific classifications he desires. If he be interested in the age distribution of all males in a given State, he will find the complete cross-classification of age and sex for each State on a separate page, and the States arranged alphabetically.

As contrasted with this general-purpose table covering one hundred and sixteen consecutive pages of a large Census volume, a special-purpose, or analysis, table will ordinarily cover not more than a page, and a reasonable rule to follow

would be that it should *never* cover more than a page. The Census volumes are usually not a good source from which to obtain illustrations of these special-purpose tables, for while they do frequently contain summary tables, these are only summaries and not in a proper sense analysis tables, since the essential purpose of the Census is to give detail and to make it readily accessible.

The construction of statistical tables may then be considered with particular reference to the special-purpose table, and following this, note may be taken of those features of construction in which the two types present contrasts.

Logical requirements for presentation in the special-purpose table. The presentation of the data in a table may be viewed as a problem of logical arrangements or logical relationships, and the mechanical features of the construction selected with a view to meeting these logical requirements. The first important requirement is proper definition. It is necessary to define or delimit the situation to be presented in the table. The table and its contents are a response to a question that has been asked, explicitly or implicitly, and the title of the table shall indicate the character of this response. One ordinarily thinks of the title as merely indicative of the contents, as logically subsequent to the contents; proper consideration of the table and of its relationship to an anterior problem requiring solution with the aid of quantitative data will place this relationship in the reverse order. The title delimits the quantitative basis of the answer sought and controls the contents. The contents follow the title. Associated with the title shall be a statement of the units in which the data are expressed and of the source from which they came, in case the materials are taken from previously published sources. See, for example, tables 8 and 9 on pages 54 and 55; also tables 13 and 20, on pages 120 and 151.

The contents of the table having thus been fixed in relation to the problem in hand, display of the relationships obtained

through classification and tabulation is the next feature of presentation in the order of importance. The construction of the table shall facilitate the placing of these materials in their proper setting; i.e., shall assist in their correct interpretation. A given classification involves a grand total or aggregate and its subdivision into a set of items coördinate with one another, the individual class items or aggregates; sometimes there are included sub-totals, or class items of a higher order of importance than the first named. Thus the population of the United States may be shown for the individual States and for certain customary groupings of States, such as New England, Middle Atlantic, etc. The items of the lowest order of importance here are population aggregates for individual States; district populations are of an order of importance one step higher, and the total stands as of the highest order. One of the most significant uses of the statistical table will have been demonstrated if it assists in presenting these various items of a classification, each on its proper level — making the individual items appear as of least importance, the sub-totals on a higher level, and the grand totals at the head of the list. This at least is the task to be imposed on the table-form; it remains to be seen if it can be performed with success. The logical problem of table structure, at this point, may be expressed as the problem of showing these coördinate and subordinate relationships.

But the logical presentation of a classification involves more than the differentiation of items from totals and sub-totals. It involves equally (1) the order of the coördinate items one with another and (2) the position of totals with reference to subordinate items. As to the location of totals, consider a classification placed in a vertical column of a table. The total may appear at the top or at the bottom, and the basis of choice between these two positions involves mainly one consideration, sometimes a second: (a) the logical relationship between total and items or (b) the reaction of the user of the table to the posi-

tion of the total. If the conception of the relationship is that of an aggregate subdivided into its parts, then the logical point of departure is the total and it should appear at the top, with its subdivisions beneath; if, on the other hand, the relationship is thought of as a series of items which, being added together, produce a total or aggregate, then the total is derived from the items and logically appears at the foot of the column. The distinction lies entirely in the point of emphasis in the relationship of subdivision to total. In table 8, page 54, the total is located at the top of the column.

Sometimes the reaction of the user of the table will control, and then this distinction may have to be abandoned. Since the purpose of the table is to aid in understanding its contents, it may be necessary to take account in its form of personal idiosyncrasies, and if a particular user for whom the table is especially made insists on seeing totals at the foot of columns, it is proper for the statistician to put the totals in this position. Though this is opportunist advice, it is such only in a specific situation and is directed to the realization of the main purpose of table structure — to aid in the use and understanding of the figures. The rule to follow in such a situation is to fit the arrangement to the psychology of the user, and, if he wants the total at the bottom, put it there. Amusing instances have occurred where tables prepared by statisticians with an eye to logical requirements failed of their purpose largely because of the prejudices of users against seeing totals at the top of the column. The statistician who learns of such a situation will unblushingly change his procedure until this prejudice has been replaced by a clearer understanding of the point at issue. In actual practice, both of the *total* positions are used, but brief consideration of cases will indicate that the differing practice reflects the logical difference in point of view here stated.

Where a classification appears in a horizontal position in the table, there is a similar choice for the *total* position at the left

or at the right of the row. The same considerations which justify the *top* in the column will select the position at the left of the row; and similarly for the foot of the column and the right of the row. See tables 8 and 9, on page 54 and 55.

For the arrangement of the sequence of coördinate items of a classification in column or in row there are likewise certain general criteria. A given classification may be any one of the four types,[1] quantitative, qualitative, geographical, or chronological. In some of these cases the bases of choice are restricted while in others several alternatives call for attention. The classes of a quantitative classification are always listed in the order of magnitude. Thus, age groups of the population:

0– 4
5– 9
10–14
15–19
20–44
45 and over.

Such an arrangement calls for no comment. If a time classification is under consideration, the classes must follow a regular order of time sequence; but this may involve, in some instances, beginning the sequence with the period latest in time, in other cases, with the earliest period. Census Bureau practice usually follows the plan, in reporting yearly figures, of presenting the latest date first, on the ground that the main interest lies in the more recent figures and that earlier figures are of interest chiefly in comparison with the latest figures; monthly figures, however, are given in a sequence beginning with January and ending with December. The interest here is in growth or development through the year.

When the classification is geographical, there is no natural method always to follow in the arrangement of classes as in the

[1] See page 7.

two previous types. In some instances there may be a more or less evident order of geographical progression such as is found in the Census classification of States. Where data are arrayed for the forty-eight States as at present, the usual order is New England States, Middle Atlantic; South Atlantic, etc.; that is, beginning on the east coast and proceeding from north to south; then the next tier of States to the west and again from north to south, and so on to the Pacific Coast. This is a reasonably familiar order to most of us. At other times the controlling motive in the geographical classification is likely to be less a matter of contiguity of area than a matter of accessibility or ready reference to the individual items, and here the alphabetical arrangement suggests itself. In other cases the purpose of the analysis is better served by arranging the classes in the order of their size. A business firm showing total sales for each of several sales districts is frequently most interested in the latter arrangement.

For a qualitative classification the arrangement by magnitude or alphabetically is frequently chosen and for reasons similar to those given for a geographical series — for comparison of size of classes or for ease of access. But this statement does not reveal all the complexities of arrangement that may arise in qualitative classifications. Simple cases, like the five or six subdivisions in the Census classification of "marital condition," do not show these complexities, nor would the ten, fifteen, or twenty coördinate *departments* of a retail department store; but the situation arises in large qualitative classifications, showing many sub-totals of various orders as well as individual items. It is difficult to give any general rule which will suffice for all cases and it is probably better to settle each complex case on its own merits. The Census classification of occupations, for instance, gives occupational designations within an industry framework, and while the general rule is for coördinate items in a given subdivision to be arranged alphabetically, it

may be necessary in order, for instance, to obtain a total of all enumerated "clerks," to search through many *industry* subdivisions to find them. This matter goes back to the difficulty, heretofore noted, of making a thoroughly satisfactory classification of occupations.

There arises a further question of importance in the proper display of relationships in a statistical table. When cross-classified data are being presented, in a double tabulation or one still more complicated, the question arises, which classification will be put in the stub and which in the caption. When the analysis, or special-purpose, table is in mind — which has been the especial object of the rules of arrangement so far given — there is a logical basis for choice. It is generally the case that one of the classifications involved is of more importance in the situation than the others, or is primary to them; and it is the general rule to place this classification in the stub, the reason being that its primacy justifies giving it the most prominent position in the table, this position being the stub, since relationships and comparisons of figures are ordinarily seen better in a vertical than in a horizontal arrangement.

Thus, in a cross-classification of sales by salesmen and by kinds of goods, if the comparison of sales of different kinds of goods is to be given the greater emphasis, this classification will be placed in the stub and the salesmen classification in the caption; while if greater importance in another situation is placed upon comparison of the results of the different salesmen, the reverse arrangement will be preferred. Sometimes this rule must be disregarded when there are many items in one classification and but few in the other, the classification containing the larger number of items being placed in the stub.

There is one final suggestion. Since analysis and interpretation is the goal sought, too much attention to insignificant detail can defeat the purpose in mind and it is often desirable to suppress such details. In comparing wheat crops of the

United States of different years, for instance, differences of one hundred thousand bushels mean little. It may suit the end in view better to express the figures, therefore, in units of million bushels or in millions significant to one decimal place; that is, units of one hundred thousand bushels. In many cases nothing more is needed than to express the figures in the form of percentages or relative numbers. When the figures involve a time classification, it is very frequent practice to select one date as the base, call this figure 100, and express the others as relative to it. This makes the comparison of different dates accurate to two or three digits and that is all that is required. In presenting department store sales, it is often sufficient to know the percentage of total sales by departments, these percentages expressed to the nearest unit. Sometimes, of course, it will be desirable to include in the table both the absolute figures rounded as above and percentages or relatives.

Modifications of these requirements for general-purpose tables. Much that has been said about appropriate methods of displaying relationships in special-purpose tables, is applicable directly to repository or general-purpose tables, but the distinction between the two types is a fundamental one and it intrudes at times to force differences in the arrangement of material. The necessity of distinguishing between relationships which are coördinate and those which are subordinate are equally great in the two cases; but whereas the totals in analytical tables are sometimes in one position, sometimes in another, according to the point of emphasis, and indeed may be omitted altogether if unimportant in the analysis and interpretation of the data, the position of totals in a repository table is likely to be wholly a matter of convenience — the point of view being no more than to place the figures where they are accessible. Accessibility, in fact, might be said to be the chief consideration in the general table.

Again, the order of the items in column or row is likely to be

determined by the same considerations in the two cases; but when it comes to choosing between column and row for a given classification in a cross-tabulation, the bases for choice for the two types of tables are distinct. Whereas in the analytical table the most important classification, or the one emphasized as primary, takes the vertical position, in the case of the general table the sole consideration is space. Accessibility is aided by compactness and the greater space is in the column; hence it is the rule in the general table always to place the longer classification in the stub or vertical position.

In the matter of rounding the figures there is also a difference in practice, in the special table the motive being to avoid unimportant detail and to use only significant digits. But the general table is conceived of as the repository of the full results of tabulation, and it is not the place for lopping off the figures. There is no thought, for instance, that the publication of the United States population total in a Census volume, giving the figures in its millions, thousands, hundreds, and units, involves any implication that the population has been counted correctly to a man, but only that the figure given represents the full final result of the count as taken; and it is the task of the Census authorities or of the user of the figure to decide within what limits the figure given can be accepted as accurate. The full results of tabulation are placed in the general table, and the rounding-off process comes later when the figures are put to special uses and estimates of their significance become important.

Mechanical aids to good tabular presentation. The proper arrangement of materials in the statistical table to display the relationships desired is, of course, of prime importance. But this part of tabular presentation may be well done and still the statistical table fail of its purpose. For not only must the data be associated part with part in accordance with logical demands, but the mechanical features of the table must be such as to aid in the portrayal of these associations.

TABLE 2.

PRODUCTION OF SIX IMPORTANT CROPS IN MINNESOTA YEARS 1924 TO 1927

UNIT: 1000 bu.

CROP	1924	1925	1926	1927
Wheat	37,863	30,269	24,811	21,397 *
Oats	197,241	200,340	129,162	120,493 *
Potatoes	44,880	26,772	29,800	33,128 *
Corn	124,065	148,896	147,662	127,246 *
Rye	14,718	5,824	5,940	7,485 *
Flaxseed	8,117	7,400	7,652	7,343 *

*Preliminary. Source: *U.S. Statistical Abstract.*

TABLE 3.

PRODUCTION OF SIX IMPORTANT CROPS IN MINNESOTA YEARS 1924 TO 1927

UNIT: 1000 bu.

CROP	1924	1925	1926	1927
Wheat	37,863	30,269	24,811	21,397 *
Oats	197,241	200,340	129,162	120,493 *
Potatoes	44,880	26,772	29,800	33,128 *
Corn	124,065	148,896	147,662	127,246 *
Rye	14,718	5,824	5,940	7,485 *
Flaxseed	8,117	7,400	7,652	7,343 *

*Preliminary.

Source: *U.S. Statistical Abstract.*

Of first importance among such aids is the correct setting of the table on the page. Every statistical table has a definitely circumscribed amount of space allotted to it; it may occupy a page or part of a page, may be associated with textual matter or not, or it may be displayed on a wall chart; but in each instance there is a given amount of space that is devoted specifically to the table and to nothing else. The mistake of most beginners in the construction of statistical tables is in thinking that this entire space must be filled with the figures, and they will therefore draw boundary lines for the table at the very limits of this space. No greater mistake could be made. With a given space, say a page, no inconsiderable element in the success of the table arises from making the table center in this space with a generous margin of white paper on sides, top, and bottom. This border of unused space serves to focus the attention at the proper place — upon the space utilized by the figures; whereas if the entire page is taken up with figures, it leaves an impression of confusion. It cannot be too greatly emphasized that the success of any statistical table is measured in terms of the reactions of those who use it, not of those who make it; and if the user is confused, or is repelled by the setting of the table on the page, this fact detracts from its success as an aid in interpretation. The effective use of margins is most clearly seen in the case of tables used for display purposes on wall charts or in tables which occupy a full page of a book or manuscript. Where they occupy only part of a page as an adjunct to printed matter, they are likely to have the same side margins as the printed matter, and no margins at all at top and bottom. These cases, however, fail to show the full possibilities of table-form as an aid in interpreting statistical materials.

The proper procedure in constructing a table is to decide on this margin first of all and thus set the limits of space to be occupied by the figures, then to make the figures of such size that they can be contained within their space allotment. One's

inclination is, of course, to look at the figures first and to arrange margins in accordance with the quantity of data to be presented. The margins should be of essentially the same size on all sides of the table; that is, the table should be centered in the allotted space. Just how large these margins should be is, again, a matter to be decided by the user. It is a question what margin gives the best impression of the table as a whole. Rigid rules are not applicable here, but there will probably be general agreement in the choice of alternatives presented in tables 2 and 3, where the sole difference lies in the margins.

The question whether the table should be boxed in by

TABLE 4.

SWEDEN'S FOREIGN SHIPPING

Totals of net tonnage entered and cleared, expressed in Index Numbers

Average for 1913 = 100

YEAR AND MONTH	TONNAGE ENTERED				TONNAGE CLEARED				TOTAL TONNAGE
	Swedish		Foreign		Swedish		Foreign		
	with cargo	in ballast	with cargo	in ballast	with cargo	in ballast	with cargo	in ballast	
1927 Maj	115	118	124	140	109	141	104	209	117
Jun.	112	153	116	233	121	141	138	258	135
Jul.	109	213	119	261	136	155	154	293	148
Aug.	115	191	144	352	122	158	171	485	153
Sep.	114	191	123	245	131	141	153	308	146
Okt.	120	160	122	211	120	136	137	259	136
Nov.	114	131	116	156	111	125	113	270	121
Dec.	103	107	109	94	102	133	101	202	106
1928 Jan.	91	37	96	43	59	145	71	226	79
Feb.	78	33	85	54	63	159	66	180	74
Mar.	82	59	101	56	73	110	82	193	84
Apr.	88	81	99	56	76	112	73	220	84
Maj	108	115	114	117	107	190	90	301	112

Reproduced from table published by Svenska Handelsbanken, Stockholm.

TABLE 5.

IMPORTS AND EXPORTS OF GUAM

YEAR ENDED	MERCHANDISE IMPORTS			MERCHANDISE EXPORTS		
	From United States	From Other Countries	Total	To United States	To Other Countries	Total
June 30:	Dollars	Dollars	Dollars	Dollars	Dollars	Dollars
1916	177,163	79,785	256,948	33,306	29,007	62,313
1917	114,301	172,351	286,652	46,972	33,363	80,435
1918	221,241	136,906	358,147	68,742	63,016	131,758
Dec. 31:						
1918 (6 months)	108,460	71,543	180,003	2,901	36,059	38,860
1919	308,465	138,716	447,181	49,222	15,330	64,552
1920	234,960	120,692	355,652	28,432	22,066	50,498
1921	304,111	179,573	483,684	15,566	24,776	40,342
1922	424,411	171,709	596,120	49,426	13,505	62,931
1923	456,824	217,732	674,556	77,109	16,977	94,086
1924	380,506	252,215	632,721	55,192	10,903	66,095
1925	324,619	261,216	585,835	89,219	10,735	99,954
1926	306,194	218,125	524,319	86,298	32,730	119,028
1927	222,818	195,617	418,435	183,579	41,408	224,987

Source: Returns to the Navy Department. Table reproduced from *U.S. Statistical Abstract*, 1928, p. 562.

boundary lines on all sides or should be left with open ends is one to which altogether too much attention may be devoted. The practice of the Federal Census Bureau is, in general, to leave the ends of the table open, and consideration of the expense of including the end lines is probably as large a factor in setting this practice as any other; but official bureaus seem to be about equally divided on the use of open ends and completely boxed tables. These two alternatives are illustrated in tables 4 and 5.

The title of the table should occupy a central position at the top of the table. Where it is brief, it may be placed upon a single line, but if too long for this, the more significant portions should be placed on the first line and the remainder below, also centered. It is usual when the title occupies more than one line to include on the first line those portions indicating the field covered by the data, or the characteristics possessed by every unit of observation, reserving for second or subsidiary lines the statement of classifications into which the data are distributed.

In some cases the first line may indicate the main classification, the remaining classifications being given in smaller type on the second line. A few illustrations follow:

(1)

FULL TIME STUDENTS REGISTERED IN ARTS COLLEGE, FALL TERM 1928

Classified by sex and by class in college.

(2)

PERSONAL AND CORPORATION INCOME TAX RETURNS

By States and Territories, 1924.

(3)

DENSITY OF POPULATION PER SQUARE MILE

Classified by States and Population Districts for Various Census Periods.

(4)

AGE DISTRIBUTION, CONTINENTAL UNITED STATES

By Classes, 1920, with Certain Comparisons for Previous Censuses.

The larger type used in the top line suggests its primary importance with reference to the materials contained in the table. In repository tables, such as the Census volumes or those in the United States *Statistical Abstract*, this method of centering and arranging the title is not followed rigidly because of the prime need for saving space. Tables 2 and 3 show well-arranged titles, the primary classifications being given on the first line and the secondary on the second. Table 6, reproduced here, illustrates a defective title, for the title does not properly show the contents of the table. The table being an integral part of a discussion of the Census of Manufactures, that fact need not

TABLE 6.

SIZE OF ESTABLISHMENT BY VALUE OF PRODUCT — 1925

SIZE OF ESTABLISHMENT	ESTABLISHMENTS		WAGE EARNERS		VALUE OF PRODUCTS	
	Number	Per cent	Number	Per cent	Amount (000 omitted)	Per cent
All classes	187,390	100.0	8,384,261	100.0	$62,713,714	100.0
$5,000 to $20,000	55,876	29.8	156,373	1.8	628,373	1.0
$20,000 to $100,000	68,951	36.8	660,309	7.9	3,272,197	5.2
$5,000 to $100,000	124,827	66.6	816,682	9.7	3,910,570	6.2
$100,000 to $500,000	42,209	22.5	1,675,911	20.0	9,576,090	15.3
$500,000 to $1,000,000	9,771	5.2	1,131,439	13.5	6,870,112	10.9
$1,000,000 and over	10,583	5.7	4,760,229	56.8	42,366,941	67.6
$500,000 and over	20,354	10.9	5,891,668	70.3	49,237,053	78.5

Reproduced from *Jour. Amer. Stat. Assn.*, Suppl. Mch. 1928, p. 30.

be given in the title, but the data included in the table are better indicated by some such title as the following:

ESTABLISHMENTS, WAGE EARNERS AND VALUE OF PRODUCTS, 1925
Classified by Size of Establishment.

It is worth while to note that the data included in table 6 involve three separate single tabulations, the data of three distinct though related fields of observation, and that the table therefore does not possess unity in the sense, sometimes employed, of showing the data of a single universe.

When the unit of measurement is stated in connection with the title, it is usually given as in tables 2 and 3. Sometimes, as in table 7, the units are shown at the head of each column, or, in case the categories of a stub classification refer to different kinds of units, a separate column may be added for the units designation; for example:

Commodity	Unit	Quantity Marketed
Wheat	bu.	
Cattle	cwt.	
Hay	tons	
Eggs	doz.	

TABLE 7.

VESSELS OF U.S.NAVY FIT FOR SERVICE, INCLUDING THOSE UNDER REPAIR: NUMBER AND DISPLACEMENT, JUNE 30

JUNE 30 —	TOTAL		FIGHTING SHIPS		NONFIGHTING SHIPS	
	Number	Displacement	Number	Displacement	Number	Displacement
		Tons		Tons		Tons
1906	276	692,592	200	518,115	76	174,477
1910	308	1,075,407	220	828,695	88	246,712
1915	343	1,352,135	230	913,334	113	438,801
1920	795	2,111,457	618	1,369,880	177	741,577
1923	774	2,353,660	585	1,333,065	189	1,020,595
1924	753	2,258,843	565	1,253,182	188	1,005,661
1925	754	2,274,376	567	1,269,791	187	1,004,585
1926	734	2,247,955	557	1,273,550	177	974,405

Reproduced from *U.S. Statistical Atlas*, 1926, p. 146.

The source of the material, in case the table does not represent an original tabulation by its maker, should always be indicated. It may be put in an inconspicuous place — just above the caption classification at the extreme right or just below the table at either side are good locations. Whenever any irregularity occurs in the data, or where there is any restriction or limitation applying to individual items and not to the entire table, these facts should be noted by starring or otherwise marking the appropriate figures and giving the explanation in a footnote. See, for example, tables 2, 3, and 9.

Column and row headings should state clearly and briefly the nature of the data to which they refer and should always be written vertically. To have to turn the table on its side in order properly to read the headings in the caption detracts from the success of the table. Sometimes columns or rows are numbered to facilitate reference, especially in the general-purpose table.

A very important element in good table-form lies in those features of construction that assist in displaying the coördinate and subordinate relationships of the various classifications. A

Illustrations of Arrangement of Stub Classifications and of Type Used

(1)

DIVISION AND STATE	Census year.	NUMBER Total.
United States	1919	**290,105**
	1914	**275,791**
GEOGRAPHIC DIVISIONS:		
New England	1919	25,528
	1914	25,193
Middle Atlantic	1919	88,360
	1914	85,466
East North Central	1919	61,332
	1914	59,896
West North Central	1919	29,166
	1914	27,199
South Atlantic	1919	29,976
	1914	28,925
East South Central	1919	14,655
	1914	14,410
West South Central	1919	13,909

Source: *U.S. Census of Manufactures*, 1920, vol. 8, Table 23, p. 102.

(2)

STATE.	Population.
United States	**105,710,620**
New York	10,385,227
Pennsylvania	8,720,017
Illinois	6,485,280
Ohio	5,759,394
Massachusetts	3,852,356
New Jersey	3,155,900
Michigan	3,668,412
California	3,426,861
Indiana	2,930,390
Wisconsin	2,632,067
Missouri	3,404,055
Connecticut	1,380,631
Minnesota	2,387,125
Texas	4,663,228
North Carolina	2,559,123
Kansas	1,769,257
Maryland	1,449,661
Washington	1,356,621

Source: *U.S. Census of Manufactures*, 1920, vol. 8, Table 10, p. 18.

(3)

DIVISION, STATE, AND SEX.	Average number employed during year.
UNITED STATES:	
1919	9,096,372
Males	7,267,030
Females	1,829,342
1914	7,036,246
1909	6,615,046
GEOGRAPHIC DIVISIONS.	
NEW ENGLAND	1,351,389
Males	955,597
Females	395,792
MIDDLE ATLANTIC	2,872,653
Males	2,179,258
Females	693,395
EAST NORTH CENTRAL	2,396,618
Males	2,030,024
Females	366,594
WEST NORTH CENTRAL	499,635
Males	408,369
Females	91,266
SOUTH ATLANTIC	817,212
Males	658,092

Source: *U.S. Census of Manufactures*, 1920, vol. 8, Table 15, p. 37.

Illustrations of Arrangement of Stub Classifications and of Type Used

(4)

Class	Pe Total
1920	
URBAN POPULATION	
Males, total	**19,695,500**
White	18,291,363
Negro	1,325,398
All other	78,739
Native white—	
Native parentage	8,350,138
Foreign or mixed parentage	4,578,547
Foreign-born white	5,362,678
Females, total	**19,618,764**
White	18,214,266
Negro	1,383,150
All other	21,348
Native white—	
Native parentage	8,547,716
Foreign or mixed parentage	5,066,800
Foreign-born white	4,599,750
RURAL POPULATION	
Males, total	**17,225,163**
White	15,044,223
Negro	2,067,813
All other	113,127
Native white—	
Native parentage	10,741,969
Foreign or mixed parentage	2,412,393
Foreign-born white	1,889,861

Reproduced from Table 37, p. 42, *U.S. Statistical Abstract*, 1928.

(5)

Class	Pe Total
1920	
URBAN POPULATION	
Males, total	**19,695,500**
White, total	18,291,363
Native white—	
Native parentage	8,350,138
Foreign or mixed parentage	4,578,547
Foreign-born white	5,362,678
Negro	1,325,398
All other	78,739
Females, total	**19,618,764**
White, total	18,214,266
Native white—	
Native parentage	8,547,716
Foreign or mixed parentage	5,066,800
Foreign-born white	4,599,750
Negro	1,383,150
All other	21,348
RURAL POPULATION	
Males, total	**17,225,163**
White, total	15,044,223
Native white—	
Native parentage	10,741,969
Foreign or mixed parentage	2,412,393
Foreign-born white	1,889,861
Negro	2,067,813
All other	113,127

A preferable arrangement of the items in (4).

given classification having been assigned to the stub, the relative significance of various categories can be shown by variations in indentation, type, and spacing. With totals, sub-totals, and individual class items to provide for, the totals may occupy a central position in the stub and be printed in capitals; the sub-totals, in heavy type, have a position well to the left, and the individual items of the classification may be printed in light type and indented. The individual items may be spaced at a constant distance apart, with greater space between them and sub-totals and a still greater space for the grand total. When there is a large number of coördinate items in the stub classification, such as the list of forty-eight States, it is helpful to provide a double space after every fifth item. This assists greatly in locating individual items in the table. These practices are well illustrated by the Federal Census classifications, several examples from which are given on pages 48 and 49.

When it comes to the subdivision of the caption classification the necessary distinctions between coördinate items and the setting-off of sub-totals and totals are accomplished by boxing arrangements, a sort of pyramid arrangement where coördinate items appear on the same level and where light lines separate columns of equal importance and heavier lines subdivisions of greater significance. The same distinctions as to type may be used here as in the stub. The caption heading used on one of the general tables of the *Census of Manufactures*, Number 53, running across two full pages of a quarto volume furnishes an illustration of this method of boxing in considerable complexity. Instead of using lines of different width, the Census volume uses single and double lines, but they are not always used in a way to maintain with consistency the distinctions between subordinate and coördinate items; and the type used is subject to the same defect. To be sure, the table in question is a general table where chief importance attaches to accessibility, and the standards here described for construction of the caption have

their principal application in the analysis table; but it may be suggested, nevertheless, that the maintenance of the same distinctions by width of line as is shown by the boxing would add to the clearness of distinctions which it is desired to show, even with the general table. In the caption in question the categories of highest order, all coördinate, each representing a main interest in the manufacturing census, and each being, in technical language, a separate field of inquiry (i.e., they do not represent subdivisions of a classification, but rather several different aggregates, some of which are divided into subclasses, and others are of interest as they stand), are the following:

1. Number of establishments
2. Persons engaged in manufacturing industries
3. Capital
4. Expenses
5. Value of products
6. Value added by manufacture
   (This is really a subdivision under item 5.)
7. Primary horsepower

It would seem appropriate to separate these various headings by heavy or double lines and to separate the subdivisions of each of these aggregates by lines of greater width than for subdivisions still further removed. This caption is redrawn on page 53 according to the suggestions made herewith. The full-page tables, numbers 8, 9, and 10, offer examples of a triple tabulation embodying all these standard features of construction.

MANUFACTURES.

TABLE 53.—DETAILED STATEMENT FOR THE UNITED STATES,

DIVISION AND STATE.	Number of establishments.	PERSONS ENGAGED IN MANUFACTURING INDUSTRIES.													
		Total.	Proprietors and officials.					Clerks, etc.		Average number of wage earners.					Wage earners employed 15th day of—	
			Total.		Proprietors and firm members.	Salaried officers of corporations.	Superintendents and managers.	Male.	Female.	Total.	16 years of age and over.		Under 16 years of age.		Maximum month.	Minimum month.
			Male.	Female.							Male.	Female.	Male.	Female.		

GENERAL TABLES.

BY GEOGRAPHIC DIVISIONS AND STATES: 1919.

Capital.	EXPENSES.								Value of products.	Value added by manufacture.	Primary horsepower.
	Salaries and wages.			For contract work.	Rent and taxes.		For materials.					
	Officials.	Clerks, etc.	Wage earners.		Rent of factory.	Taxes, Federal, state, county, and local.	Total	Principal materials.	Fuel and rent of power.			

CAPTION HEADING OF A GENERAL CENSUS TABLE

[*Census of Manufactures*, 1920, Vol. 8.]

Note arrangement of boxing to show coördinate and subordinate relationships. Use of single and double lines fails to maintain these distinctions.

TABLE 53.—DETAILED STATEMENT FOR THE UNITED STATES

DIVISION AND STATE	NUMBER OF ESTABLISHMENTS	PERSONS ENGAGED IN MANUFACTURING INDUSTRIES														
		Total	Proprietors and Officials					Clerks, etc.		Average Number of Wage Earners					Wage Earners Employed 15th Day of —	
			Total		Proprietors and Firm Members	Salaried Officers of Corporations	Superintendents and Managers	Male	Female	Total	16 years of age and over		Under 16 years of age		Maximum Month	Minimum Month
			Male	Female							Male	Female	Male	Female		

BY GEOGRAPHIC DIVISIONS AND STATES: 1919.

CAPITAL	EXPENSES									VALUE OF PRODUCTS	VALUE ADDED BY MANUFACTURE	PRIMARY HORSEPOWER
	Salaries and Wages			For Contract Work	Rent and Taxes		For Materials					
	Officials	Clerks, etc.	Wage Earners		Rent of Factory	Taxes, Federal, State, County, and Local	Total	Principal Materials	Fuel and Rent of Power			

CAPTION HEADING OF PAGE 52 REDRAWN

Width of line here reinforces the boxing arrangement to show coördinate and subordinate relationships.

TABLE 8.

NUMBER OF FARMS OPERATED IN THE SOUTHERN STATES, 1925

BY COLOR AND TENURE OF FARMERS

DIVISION AND STATE	TOTAL		BY OWNERS AND MANAGERS		BY TENANTS	
	White Farmers	Colored Farmers	White Farmers	Colored Farmers	White Farmers	Colored Farmers
THE SOUTH						
SOUTH ATLANTIC						
Delaware						
Maryland						
District of Columbia						
Virginia						
West Virginia						
North Carolina						
South Carolina						
Georgia						
Florida						
EAST SOUTH CENTRAL						
Kentucky						
Tennessee						
Alabama						
Mississippi						
WEST SOUTH CENTRAL						
Arkansas						
Louisiana						
Oklahoma						
Texas						

Adapted from *U.S. Statistical Abstract*, 1926, Table 539.

TABLE 9.

NUMBER OF FARMS OPERATED IN THE SOUTHERN STATES, 1925

BY COLOR AND TENURE OF FARMERS

DIVISION AND STATE	TOTAL		BY OWNERS AND MANAGERS		BY TENANTS	
	White Farmers	Colored Farmers	White Farmers	Colored Farmers	White Farmers	Colored Farmers
THE SOUTH						
SOUTH ATLANTIC						
Delaware						
Maryland						
District of Columbia						
Virginia						
West Virginia						
North Carolina						
South Carolina						
Georgia						
Florida						
EAST SOUTH CENTRAL						
Kentucky						
Tennessee						
Alabama						
Mississippi						
WEST SOUTH CENTRAL						
Arkansas						
Louisiana						
Oklahoma						
Texas						

Adapted from *U.S. Statistical Abstract*, 1926, Table 539.

TABLE 10

TOTAL RESOURCES OF COMMERCIAL AND SAVINGS BANKS UNITED STATES AND OUTLYING POSSESSIONS, 1914–1926

BY CLASS OF BANK

[Unit: $1,000,000]

YEAR	ALL CLASSES	NATIONAL BANKS	STATE (Commercial) BANKS	LOAN AND TRUST COMPANIES	STOCK SAVINGS BANKS	MUTUAL SAVINGS BANKS	PRIVATE BANKS
1914	26,971.4	11,482.2	4,353.7	5,489.5	1,196.5	4,253.0	*
1915	27,804.1	11,795.7	4,399.6	5,873.1	1,238.7	4,319.4	177.7
1916	32,271.2	13,926.9	5,553.0	7,028.3	1,033.3	4,547.9	*
1917	37,126.8	16,290.4	6,799.7	7,899.8	1,127.9	4,811.0	197.9
1918	40,726.4	18,354.9	7,815.7	8,317.4	1,183.2	4,818.6	236.6
1919	47,615.4	21,234.9	11,701.6	7,960.0	1,281.3	5,171.6	266.1
1920	53,079.1	23,411.3	14,009.8	8,320.0	1,506.4	5,619.0	212.6
1921	49,671.4	20,517.9	14,199.1	8,181.1	557.9	6,040.1	175.3
1922	50,425.4	20,706.0	13,064.4	8,533.9	1,583.9	6,351.6	185.5
1923	54,034.9	21,511.8	14,162.9	9,499.3	1,790.7	6,904.8	165.5
1924	57,144.7	22,565.9	14,816.0	10,323.8	1,923.4	7,364.7	150.9
1925	62,057.0	24,350.9	15,972.2	11,565.5	2,093.1	7,913.0	155.2
1926	64,893.4	25,315.6	16,579.7	12,205.2	2,196.4	8,422.3	174.2

*Data not given. Source: *U.S. Statistical Abstract*, 1926, p. 253.

[Note. — Extended notes concerning the accuracy and completeness of the data given in the source quoted, are omitted here, since the sole interest is in the mechanical form of the table.]

EXERCISES

Note: Good source material from which to draw illustrations of different kinds of classifications and different sorts of cross-classifications will be found in the United States Census volumes and the *United States Statistical Abstract.*

I

Select from the pages of any volume of the *United States Statistical Abstract* illustrations of each of the four kinds of classification; select a double tabulation and specify the kinds of classification; similarly for a triple tabulation.

II

The following table is reproduced from the *United States Statistical Abstract*, 1928:

No. 197.—CORPORATION INCOME TAX RETURNS: CORPORATIONS DISTRIBUTED ACCORDING TO SIZE OF NET INCOME, BY INDUSTRIAL GROUPS

[All money figures in thousands of dollars]

ALL CONCERNS

Income class	Number	Net income	Number	Net income	Number	Net income	Number	Net income
	1921 [1]		1922 [1]		1923 [1]		1924 [1]	
Reporting net income	171,239	4,336,048	212,535	6,963,811	233,339	8,321,529	236,389	7,586,652
Less than $10,000	135,987	328,113	160,548	393,829	175,111	440,400	180,275	440,136
$10,000 to $100,000	29,922	867,916	43,123	1,279,204	48,022	1,468,860	47,268	1,404,772
$100,000 to $1,000,000	4,799	1,250,407	8,019	2,159,112	9,180	2,421,598	7,945	2,083,697
$1,000,000 to $5,000,000	461	918,042	725	1,444,773	858	1,695,717	739	1,447,837
$5,000,000 and over	70	971,570	120	1,681,893	168	2,294,954	162	2,210,210
Reporting deficit	185,158	[3] 3,878,219	170,348	[2] 2,193,776	165,594	[2] 2,013,554	181,032	[2] 2,223,926

1925

	Total, all concerns [1]		Agriculture		Mining and quarrying		Manufacturing total	
Reporting net income	252,334	9,583,684	4,662	76,862	5,488	453,600	54,137	4,383,357
Less than $10,000	188,848	467,453	3,805	8,451	3,718	7,775	33,420	92,436
$10,000 to $100,000	52,747	1,621,589	773	20,987	1,319	45,823	15,769	535,358
$100,000 to $1,000,000	9,626	2,520,788	78	16,917	383	105,831	4,407	1,209,085
$1,000,000 to $5,000,000	917	1,876,243	5	8,041	56	128,889	446	900,082
$5,000,000 and over	196	3,097,611	1	22,466	12	165,582	95	1,646,396
Reporting deficit	177,738	[3] 1,962,628	5,242	[3] 59,215	13,675	[3] 209,957	34,537	[3] 682,255
	Food		Textiles		Leather		Rubber	
Reporting net income	9,303	533,472	7,504	413,115	1,373	76,023	349	122,966
Less than $10,000	6,215	16,588	4,644	12,573	836	2,453	184	524
$10,000 to $100,000	2,522	80,065	2,137	73,725	419	14,357	95	3,444
$100,000 to $1,000,000	493	131,891	664	175,495	107	26,699	55	18,385
$1,000,000 to $5,000,000	56	113,050	57	110,371	10	18,105	10	20,107
$5,000,000 and over	17	191,878	2	40,951	1	14,409	5	80,506
Reporting deficit	5,419	[3] 91,512	4,767	[3] 114,772	986	[3] 28,895	289	[3] 13,941
	Lumber		Paper		Printing		Chemicals	
Reporting net income	4,657	200,316	1,288	111,186	6,523	190,909	3,951	623,277
Less than $10,000	2,573	7,962	633	2,093	4,872	12,153	2,266	5,774
$10,000 to $100,000	1,660	59,914	453	16,108	1,359	44,501	1,206	42,330
$100,000 to $1,000,000	408	103,240	180	46,497	267	71,193	410	119,332
$1,000,000 to $5,000,000	16	30,190	21	40,950	23	49,356	52	110,820

[1] Including combination enterprises and inactive concerns not shown separately in 1925. [2] Deficit.

Source: Statistics of Income, Report of the Commissioner of Internal Revenue, Treasury Department.

1. Enumerate the kinds of classification in the table according to the classification given in the text.
2. Make a list of all the single tabulations that can be made from

the above table; of all the double tabulations; of all the triple tabulations.

III

An automobile manufacturer produces cars in four models and sells in ten States. Draw up a tabulation sheet to show monthly sales (numbers of cars and value of sales in dollars) by models and by dealers in each State.

IV

(1) Construct a statistical table to compare the value of yearly sales of different dealers for each of the four models above; (2) a table for sales of each model in each State. Give reasons for your selection in each instance of classifications for stub and for caption.

V

A triple tabulation formed from the population of a State classified by sex, color, and marital status, will require that two classifications appear either in stub or in caption. Make rough sketches of the possible alternative arrangements and give reasons for selecting each of the different alternatives.

VI

A street-car corporation seeks legislation restricting the right of motor busses to operate in competition with them, claiming that the busses take so much patronage that the trolley company earnings are reduced below a legitimate return upon capital investment. Suppose the municipal authorities decide that they must act only upon a basis of pertinent facts and order a survey. Suggest classifications and cross-classifications of facts which might be valuable for this purpose and sketch the form of tables in which the data are to be placed.

VII

Select any available pamphlet, monograph, statistical report or other source of tabular statistical data and criticize the tables on the basis of the standards set in the previous pages.

PART II
GRAPHIC STATISTICS

I •

INTRODUCTORY

What graphic presentation adds to the figures. Tables 8, 9, and 10 have been constructed in accordance with the standards of good table-form described in the previous pages. A prime consideration in setting up these standards has been to enable the user to see the contents of the table as a unified body of data bearing upon some particular problem. The separate parts of the table have been important in themselves, but they have been utilized, in the table, principally through relating them, the one to the other, for the purpose of creating the unified whole. In building up standards of construction, part by part, emphasis has been placed upon the necessity of displaying the logical relationships of part to part and of part to whole. The completed table, therefore, may be a complicated set of relationships even though each classification in itself be relatively simple, especially where several classifications are combined in a triple or quadruple tabulation. The final success of the table is measured in terms of the total conception, but is dependent upon accuracy in detail. Emphasis was laid in the beginning on the fact that the table-form adds nothing to the meaning of the figures, but that its essential utility lies in the assistance it gives in getting at that meaning.

It is at this point that statistical graphs bear a similarity to statistical tables; for graphs are devices to aid in the interpretation of the figures. They should not be thought of as substitutes for the figures, and ordinarily the figures should be presented in an appropriate table along with the graphs. It is well to keep in mind at all times the fact that correctness of interpretation is the aim of all methods used in presenting

statistical materials — these methods are but aids in seeing and in understanding. Statistical graphics share with many other things the danger of being overdone when their use is being widely extended; and the last ten or fifteen years have seen an enormous growth in the use of graphic methods. This growth has been generally along sound lines, but there is still need for emphasizing fundamentals.

The fundamental methods of graphic representation; their accuracy compared. Tables 8 and 10 contain materials, the presentation of all of which in graphic form would call in aid a large proportion of the recognized standard methods of graphic representation. To illustrate, many simple comparisons may be made, such, for instance, as the resources of national banks and of private banks in any given year; the proportion of all banking resources possessed by each class of banks; the number of farms in any given area operated by white and by colored farmers; the number operated by owners and managers as compared with tenants; these same facts compared in proportionate terms. Any given category in the caption of table 10 may be shown for each year 1914 to 1925; or in table 8 each caption class may be shown in a geographical distribution. Similarly many of the data in these tables may be expressed in relative form and the relative items shown by appropriate graphic methods. For instance, the number of farms tenant-owned may be expressed in proportion to the total farms operated and these data shown graphically for each State in the South.

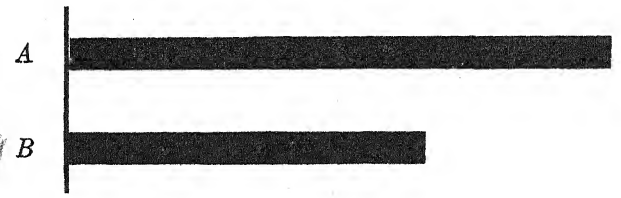
The common feature running through all these cases, as it is common of course to statistical methods as such, is that comparison is made of two or more things in quantitative terms. It is dollars of resources in national banks and in savings banks, or numbers of farmers in each Southern State; or, in other comparisons, it may be *bushels* or *tons* or *dozens* or *days* or *miles*. The items to which these varying quantities apply may repre-

sent the categories of any one of the four types of classification discussed in the previous pages; and the comparison may be of the varying magnitudes of two items or of several. If several items are involved, the purpose may go no further than to array them in the order of magnitude; if a time classification, the order of the items is predetermined and the purpose is to study variation in time; or for a geographical classification, the purpose may be either an array according to magnitude or the *distribution* of the items according to geographical location.

The fundamental problem of graphic statistics is to select from among various alternatives appropriate graphic methods of making these comparisons, and the test of success of the choice made lies in the speed and accuracy with which the comparison is made as against comparison of the figures. Graphic portrayal of magnitudes may be in terms of one-, two-, or three-dimensional space, or in terms of angular measurement. Or, concretely, magnitudes may be represented by the lengths of lines drawn to a common scale, or by areas with a common unit of area, or by volumes with a common unit of volume, or they may be represented by angles of which the common unit is the degree. The choice of alternatives is a choice among these four and the decision rests upon the question which of the four methods of comparison leads most quickly and with greatest accuracy to the meaning of the figures. Consider, for instance, the simplest sort of illustration, a comparison of two magnitudes, A and B, to determine by how much the larger exceeds the smaller. The correct answer may be, A is one third larger than B, or ten or twenty or seventy-five per cent larger. That graphic method of representing these two magnitudes is best which enables observers to estimate this ratio with the greatest accuracy. The choice of method, therefore, essentially reduces to a problem of average errors; for the estimates by different observers of any graphic method will necessarily differ, but the method that shows the

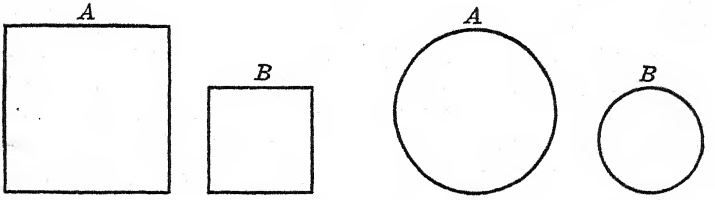
smallest difference on the average between the actual ratio and the estimated ratio is, by this test, to be preferred.

The comparison of A and B may be shown by the length of two lines drawn to a common scale; for instance:

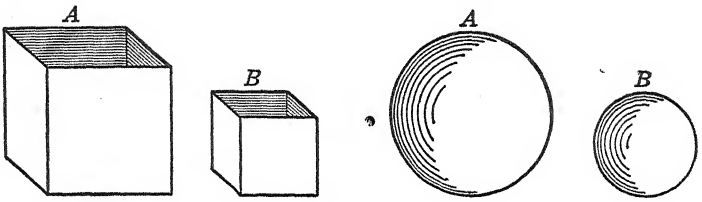


Since the bars are of equal width, they differ only in length and therefore furnish a linear comparison; and since they start from a common position at the left, the comparison is made in terms of the distance which each bar extends to the right of this common origin.

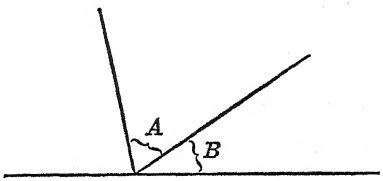
Again, the comparison may be made in terms of two-dimensional space; that is, two areas. The area comparison would be of extreme difficulty if A and B were of irregular shapes; a fairer test of this method will be obtained if the areas are squares or circles, thus:



The three-dimensional space comparison, or comparison of two solids, may be made by constructing A and B as two cubes or two spheres the volumes of which bear the ratio of A to B, thus:



Comparing the two magnitudes as angles may be done most simply as follows: The two angles, *A* and *B*, may be drawn on separate diagrams, but the comparison will then be more difficult than when the angles have a common point of origin.



These four devices, then — lines, areas, volumes, and angles — constitute the fundamental methods of graphic comparison of magnitudes. There are few statistical graphs in use to-day that do not involve use of one of the four, and the exceptions are primarily in the field of geographic comparison. The point of present interest is to choose from among these four methods the one or ones that result in most accurate interpretation of the figures. It is a question which of the devices illustrated above are most easily and most accurately compared. Is it easier to estimate the ratio of one line to another, or of the area of one square or circle to another, or of the volume of one solid to another? The question is answered almost without exception by observers that, of the three, the line comparison is most easily and accurately made. To make the area comparison with the square figures, the usual procedure would be for the observer to estimate the sides of each figure and then to compare the products of these two measurements; but this is a fairly complicated operation in any instance where the purpose is to give a rapid and accurate impression, and many people

who have use for graphs would not find this process of calculation easy. Comparison of the areas of two circles is likely to be as difficult as that of the two square areas, if not more so, for few people would know the method of calculating circular areas and the resulting estimate would be little more than guesswork. Comparison of volumes involves calculation with three dimensions and is more difficult still. A few experiments at estimating ratios of two lines, two areas and two volumes will readily bear out the conclusion that the first of the three is by far the easiest to make and the most accurate.[1]

Comparison by means of angles remains to be considered. Angle comparisons are generally made by constructing circles and then marking off a sector or sectors in distinctive fashion, so that one sector may be compared with another or with the entire circle. The circle and sector comparison is, therefore, especially appropriate when the viewpoint is comparison of a part to a whole. Actual experiments in estimating the ratios of part to part or of part to whole by this method indicate that its rating in accuracy is about the equivalent of that of linear comparisons. These two, then — linear and angular representations — are to be preferred over areal and spatial representations whenever the graphic method is utilized to make magnitude comparisons.

In summary, then, the fundamental principle of graphic representation of statistical facts lies in substituting for the figures themselves a geometric representation of them, expressed in measurement units. While two magnitudes may be compared by direct use of the figures, it is ordinarily not an easy matter to estimate their relationship with a high degree of accuracy by this means; if the figures are reduced to relative terms, ease and accuracy of comparison are increased, but this

[1] The writer has, from time to time during the last ten years, tried this experiment on classes numbering over one hundred and the results have uniformly borne out the conclusion stated above.

procedure involves calculation and therefore effort on the part of the user; and the graphic method of statistics is designed with the special purpose of getting at the meaning of the figures with a minimum of effort. When several magnitudes are to be compared, even relative figures fail to give an instantaneous picture of their interrelationships, for each relative must be noted in turn, whereas graphic methods are available that will give at a glance a picture of the entire set of relationships and with a high degree of accuracy. When it comes to the matter of choice between different measurement units for the geometric representation of the figures, both logic and experience point to the preference of linear and angular units over area- and volume-representations. These conclusions form the basis for judging the qualities of different kinds of statistical graphs considered in the following pages.

EXPERIMENTS IN GEOMETRIC REPRESENTATION OF MAGNITUDES

Various experiments may be performed with a class or other group of persons to test the comparative accuracy of different kinds of devices for graphic representation and the results of such experiments will be most valuable in checking the views and beliefs held at present with reference to the comparative values of different methods of representation. Very few of the graphic methods in use to-day have been tested widely by experiment. Their validity is founded largely on logical considerations; and while the importance of this foundation is not to be denied, no less is it to be denied that these conclusions deserve to be subjected to careful testing wherever experimental methods may be applied and where the results may be set forth in quantitative terms.

The few experiments given here are suggestive of a procedure that may be followed for testing many alternative methods of graphic representation. If such tests could be made under carefully controlled conditions by even a small percentage of the teachers of statistical methods and the results made available for all, a large body of valuable information would be obtained.

I

Linear measurements. Relative accuracy of horizontal and vertical bars for comparing two magnitudes.

(a) Construct two bars of the same width but of different lengths, drawn vertically from the same horizontal base line. Have students view the diagram for a few seconds only and have each student record on a card (1) the percentage which the shorter bar bears to the longer, or (2) the percentage which the longer bar bears to the shorter. It would be desirable to have (1) and (2) done at different times or upon different groups. Furthermore, the test should be made for several pairs of bars bearing different ratios. Bars showing 25, 50, or 75 per cent ratios may be estimated with an accuracy quite different from other ratios, such as 60, 85, 90, etc.

(b) Perform same experiment as in (a) but with bars extended horizontally from a common origin at the left. The accuracy of (a) and (b) may then be compared by constructing frequency distributions of the results and comparing means and measures of dispersion of the estimated ratios with the true ratios obtained by actual measurement.

II

Graphic comparison of two magnitudes. Comparative accuracy of lines or bars, of areas and of volumes.

(a) Construct two bars as in I (a) or I (b) of lengths in proportion to the two magnitudes to be compared. Construct also on a different diagram two area figures (squares or rectangles) and similarly on a third two volumes (cubes or spheres) to represent the two magnitudes. Then have the group estimate the ratios of the two magnitudes as in I (a) and analyze the results as in I (b).

III

Graphic comparison of a whole and its parts. Comparative accuracy of subdivided bar and of circle and sector diagram.

This experiment should be performed for a total divided into two parts of varying sizes, such as 17 per cent and 83 per cent, 42 per cent and 58 per cent, and not alone for the more obvious ratios like 50 per cent and 50 per cent, 25 per cent and 75 per cent, etc.; it

should also be performed for a total divided into several parts, say four to six or eight.

The whole and its parts should be represented graphically by a subdivided bar, and also by a circle and sector diagram, each constructed as described on pages 78–88. Estimates of the proportions of various parts to the whole may then be made by the class from each diagram and the results compared as before.

II

COMPARISONS OF MAGNITUDES AND COMPONENT PARTS — PICTOGRAMS

THE kinds of graphs used to represent statistical facts may be classed for convenience into three groups: pictograms, statistical maps, and curves. Pictograms are those graphs used for direct and simple comparison of magnitude,[1] where the point of view is either comparison of coördinate items or of parts with a total. Again, for convenience these will be referred to as (1) magnitude comparisons and (2) component parts' comparisons. With statistical maps, the purpose is always to show distribution in space of frequency or of measurable magnitude; while curves are of two general kinds: (1) frequency curves, where the purpose is to show variation in frequency with variation in magnitude: and (2) historical curves, showing variation in some quantity over time.

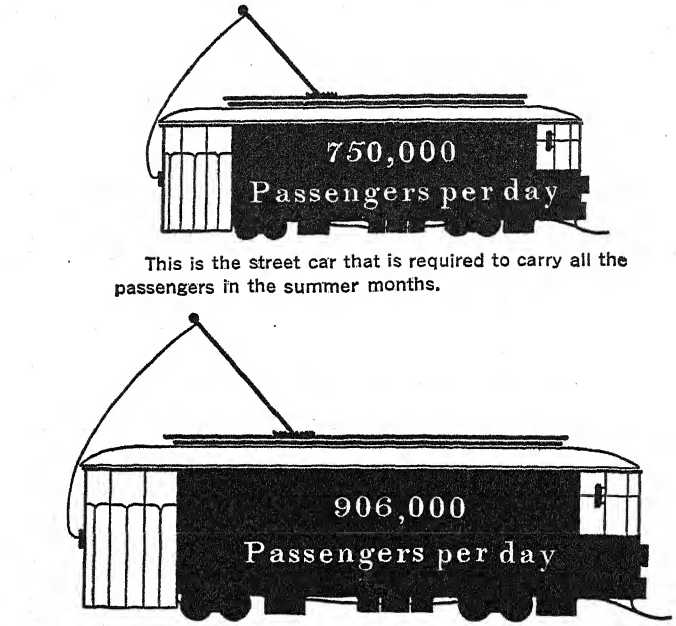
Kinds of pictograms. Of the various statistical graphs that may be classed as pictograms, the following may be enumerated:

1. Bar diagrams
2. Rectangles and circles
3. Cubes and spheres
4. Circles and sectors
5. Pictures

[1] The word *magnitude* is used here in a broad sense to include (1) those quantitative measurements arising through the use of some sort of measuring scale, such as feet or inches or days; and (2) aggregates resulting from counting or adding up individual elements where the elements or units have, or can have, an independent existence. Thus a population total is a magnitude in this broad sense, obtained from counting individual human beings. Steel production is similarly a total obtained by counting tons — the individual tons may be, but usually are not, separate entities. The distinction between (1) and (2) is sometimes difficult to make; but there are cases, in graphic statistics especially, where it is important. It is the distinction specifically between a *measurable* magnitude and *countable* frequencies. The two need not be differentiated in the discussion of graphic comparisons involved here, but the distinction is of much importance in the discussion of statistical maps. (See page 163.)

Under the last are included all representations of statistical data shown by means of actual pictures of the thing in question. The following are a few examples: (1) comparing potato crops by potatoes of different sizes, or wheat crops by wheat sacks of different sizes; (2) comparing the sizes of armies by soldiers of different sizes; or (3) showing the number of students entering different professions at two periods of time by human beings of varying sizes, each dressed to represent a given profession.

For the first four types enumerated above the basis of comparison is evident. The first is a linear measurement and comparison is made in terms of lengths of line; the second compares areas and the third volumes; while the fourth involves comparison of angles. The discussion on pages 63 to 67 has afforded a basis for choice among these four and points to the selection of linear and angular measurements alone for making these comparisons. As regards the last item in the list, actual picture representation, the fundamental criticism arises at once that it is impossible to be certain what measurement is used for comparison. It is probable, or at least possible, that the impression gained from observing pictures of two soldiers arises from a comparison of their heights, whereas the correct magnitudes may be represented either by the areas or by the cubic contents of the two figures. Similarly, the comparison of the two street cars on page 72 may be of linear dimensions, of areas or of space content. As a matter of fact, none of these measurements gives the correct comparison, though the heights or lengths of the two cars come most nearly to the correct figure, as can be shown by a few seconds' use of a ruler. The motive which leads to the frequent use of pictures in graphic representation is undoubtedly the vivid impression created by them, and it is sometimes possible to retain this vividness and at the same time produce a picture comparison that is reasonably accurate. For instance, instead of comparing two soldiers



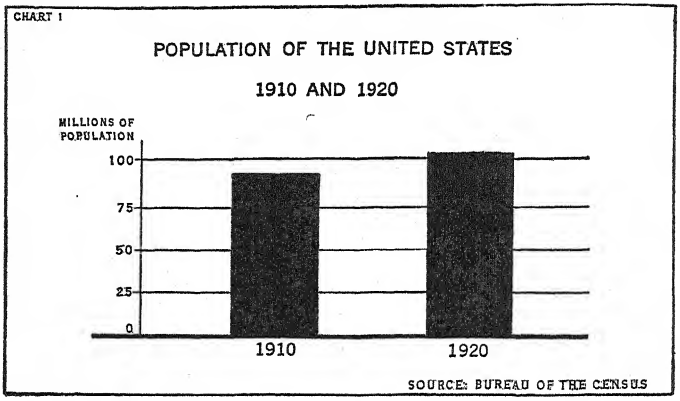
of different sizes, the two armies may be represented, each by a line of soldiers, each soldier being of the same size, but the number in each line being proportional to the number in each army. In this way the vividness of the picture representation is retained, but the measurement at the basis of comparison is linear and therefore the accuracy of the linear comparison is

realized. While by this means the picture method of graphic representation may be utilized with a considerable degree of success, it can be stated as a general rule that graphic representation of magnitudes and of component parts is best accomplished by use of the bar diagram or the circle and sector diagram. The actual cases considered below always involve one or the other of these two fundamental methods.

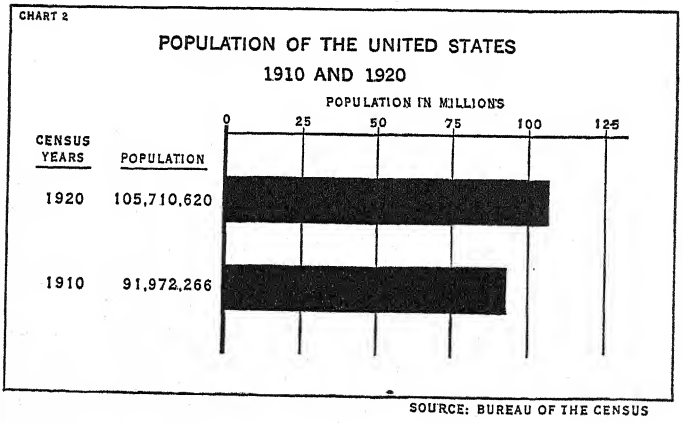
Use of pictograms — magnitude comparisons. Suppose it be desired to make a graphic comparison of two magnitudes, such as the populations of the United States in 1910 and 1920. In the general set-up of this graph on a page, as of all other graphs, what has been said about the set-up of a statistical table on a page [1] is directly applicable. Given a certain space that is to be devoted to the graph, a generous margin of white space serves to center attention where it is required — upon the graph. The title and the designation of units and sources likewise are to be drawn the same as for the statistical table. Emphasis is now laid specifically upon the construction of the graphic device which is to represent the two populations.

Since this case involves a direct comparison of two coördinate categories, namely populations for two census periods, the appropriate graph is a bar diagram, and charts 1 and 2 show two methods of constructing the diagram, with the bars drawn vertically in the one case and horizontally in the other. In each instance a scale is shown and in each instance the two bars, for 1910 and 1920, are of constant width, so that the comparison is made in terms of their variable lengths referred to the scale of population. The light scale lines on each chart aid in the comparison of the lengths of the two bars. It is the general judgment that chart 2 is to be preferred over chart 1. Experience in testing the two graphs with students indicates a general preference for the *horizontal* bar, and, other things being equal, this preference, if indicative of a general preference of all users

[1] See pages 39 to 46.



of the graph, should be determinative. There is one other important advantage of the horizontal bar graph, namely, the greater ease in including the figures; for here the space may be



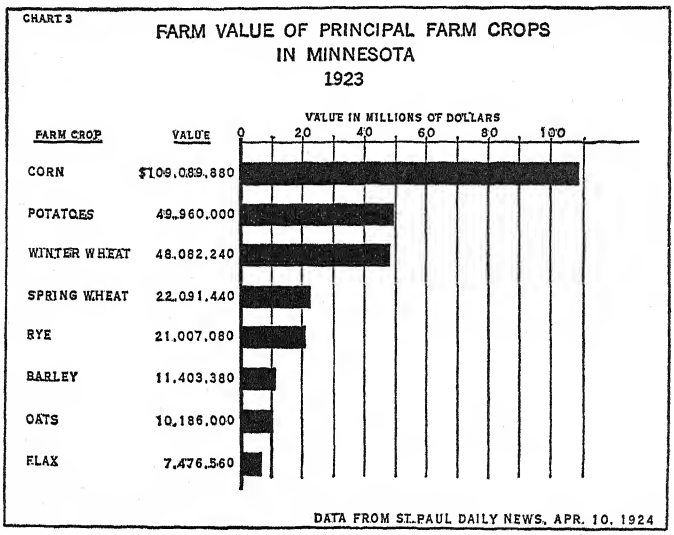
so allocated that the figures are put in a column at the left, whereas in chart 1 the figures would need to be placed in a table inset, and it is generally not so easy to find convenient space for

it. The practice of placing the figures below or above the vertical bars or within them is not in any case good usage. It is a good general rule to follow never to place figures in close juxtaposition with the bars — that is, above or below them or at the outer ends or alongside or within them — for this will detract from using the bar length and that alone as a measure of a given magnitude. Placing the figures in a vertical column at the left of the horizontal bars — that is, at the left of their common origin, the zero-line of the scale — results in no confusion of the figures as part of the magnitude portrayed, and this fact alone is a sufficient reason for choosing the horizontal in preference to the vertical arrangement.

In the construction of the scale and bars it is well to make the line drawn through the zero position on the scale heavier than the scale line, since it is the origin or point of departure of both bars. The reference lines drawn through certain points of the scale, 25, 50, 75, etc., in chart 1, should be very light lines, for they are merely supplementary aids in reading the essential facts from the bars. The bars themselves should always be of the same width and with a generous space between them, and the bars should be cross-hatched or shown in solid black, so that there can be no confusion between them and the intervening spaces. The actual width of the bars will vary with the amount of space available for the complete graph, the general rule being to use all the space available for the actual bars and to make them large enough so that they at once become the focus of attention of the observer.

Magnitude comparisons involving more than two magnitudes. Two items involve the simplest sort of graphic comparison. More frequently several coördinate items are to be shown on the graph; for example, the population of each of the forty-eight States, or the sales in each department of a business, or the value of each of several important crops. Since the purpose here is to show the importance of each item in relation to the

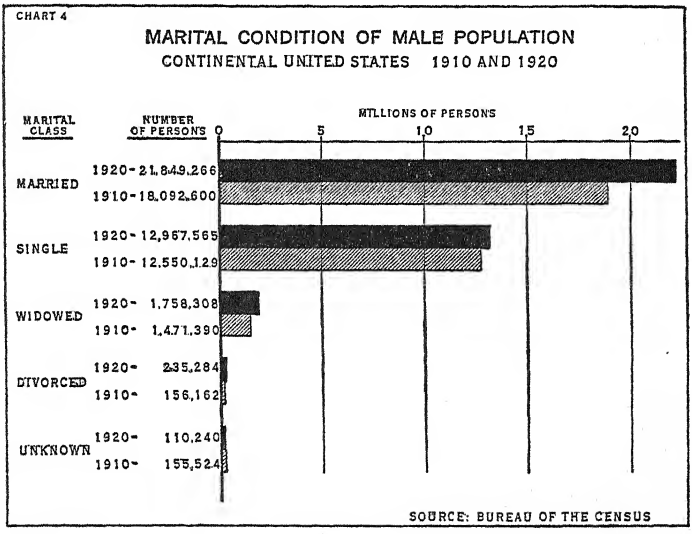
others, the bar diagram again is the appropriate method to use. It is illustrated in chart 3 showing the farm value of principal farm crops in Minnesota, 1923. This chart differs from chart 2 only in showing bars for eight categories in place of two. Attention is called to the practice of arranging the bars in the order of length with the longest at the top. While the reverse order is possible, usage generally follows the method shown



here. There is one exception, where the various categories are subdivisions of a total and for completeness an *all other* item is listed. It is correct practice to place this item at the foot of the list, since it is the aggregate of several small magnitudes which if listed separately would appear at the bottom of the series.

It is possible to devise very complicated graphs while utilizing the principle of the bar diagram for comparisons, but it is

questionable whether much is gained by the graphic method when the diagram becomes too complex. One more case may be considered as not beyond the possibilities of the graphic method. It is sometimes desired to show several magnitudes for each of two periods of time — either a geographic or a qualitative series may be shown for, say, the census periods, 1910 and 1920. Thus chart 4 shows the 1910 and 1920 census



male populations of the United States classified into marital classes. The bars are arranged in the order of magnitude of the 1920 classes. It may be noted that the emphasis in this graph is placed upon direct comparison of the various classes one with another and for one census period with another. The same figures may be used for showing the *relative* distribution of males among the various classes, but this will call for the kind of graph discussed in the next section.

Uses of pictograms for showing component parts. The relationship here designated as *component parts* involves a total and the subdivision of that total into its parts; interest is now centered upon the individual parts *in their relationship to the whole.* The previous paragraph and chart 4 refer to a case in which the figures given represent subdivisions of the total male population of the United States, but in that instance no note was taken of the total as such, comparisons being made of one part with another, each as independent entities. That use, however, did not exhaust the possibilities of the figures; situations might frequently arise in which it were desired to see the ratio of married to all males or of single to all males or to make these comparisons for each of the two census periods. When this use of the figures is in mind, a graph of the type of chart 4 is not satisfactory, for it gives no assistance in seeing the relationships of parts to whole. Other cases occur in great numbers where this latter relationship is the important one. For many geographic and qualitative series, the percentage distribution of the items of the series is one of the most significant relationships: thus, (1) the percentage sales of a store by departments; (2) the distribution of tax collections by sources or of governmental expenditures by class of expenditure; (3) the proportionate distribution of sales by sales districts or of population by states. The list could be extended with great ease, for the percentage or relative distribution of items of a statistical series is very frequently used and is of great importance.

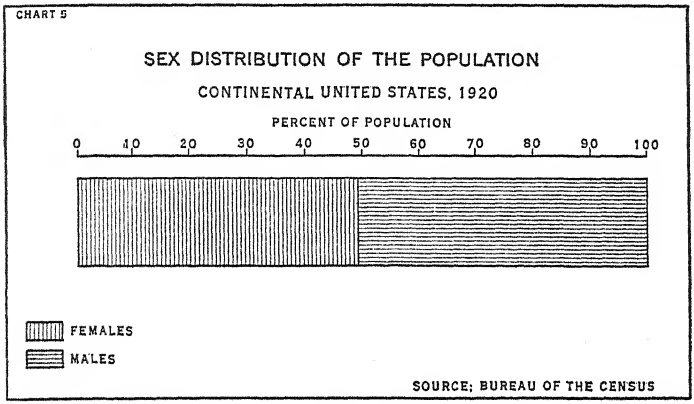
In general, there are two approved graphic methods of showing component parts: the bar diagram subdivided in the ratio of the various components; and the circle and sectors. These are to be compared and preferences determined wherever and whenever a basis for preference can be established. It will be well to illustrate first with the simplest cases and to proceed thence to the more complex.

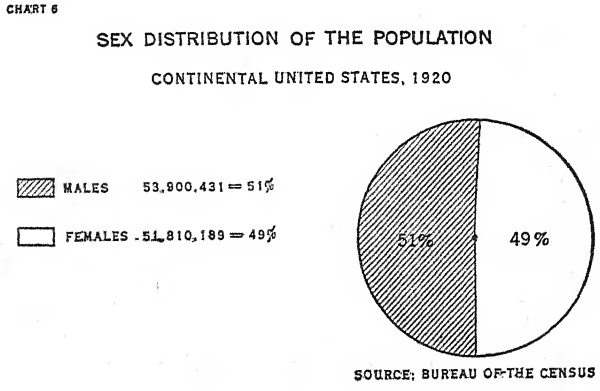
Consider then a case of two components, the population of

the United States divided into male and female. The 1920 Census figures show for continental United States:

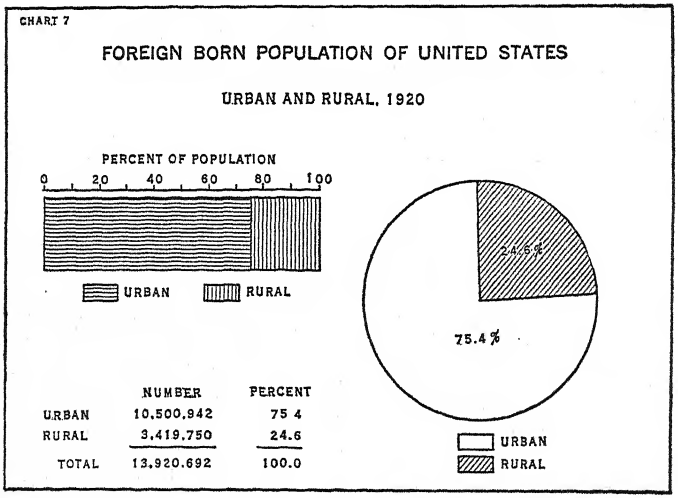
Males	53,900,431	51 per cent
Females	51,810,189	49 per cent
Total	105,710,620	100 per cent

Charts 5 and 6 show these facts graphically, one by the subdivided bar and the other by the circle and sector method.





Here is afforded a good opportunity to compare the two methods in a specific instance. Each method shows the approximately equal distribution of the sexes; but when it comes to the matter of determining which one represents the greater proportion of the total, it is quite impossible to decide this from the subdivided bar without reference to the scale above; but the circle and sector diagram gives the answer with certainty. In this case the latter scores ahead of the former. So far as concerns the general case of a total divided into two parts, this instance does not furnish assurance that a similar preference will always exist. Consider a different subdivision, that of the foreign-born population of the United States divided into urban and rural. The figures and the two graphs are shown in chart 7. Here there is probably less certainty about the result, but



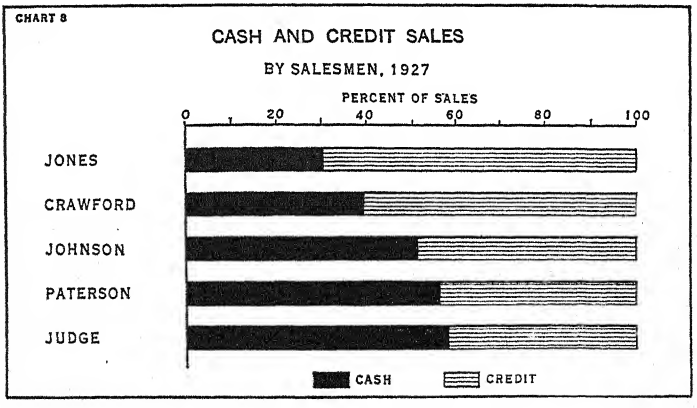
	NUMBER	PERCENT
URBAN	10,500,942	75 4
RURAL	3,419,750	24.6
TOTAL	13,920,692	100.0

SOURCE: BUREAU OF THE CENSUS

the preference seems still to lie with the circle and sector. The choice may well be still less certain for percentages differing considerably from those given in the two illustrations. When

percentage scales are included on both diagrams, they of course assist greatly in the measurement of parts and leave little basis for choosing between the two diagrams.

Sometimes it is desired to make several comparisons of totals each divided into two categories; as, for instance, showing the proportion of urban and rural in the foreign-born population of each State, or showing cash and credit sales for each of several salesmen. In such a case the bar diagram possesses a slight advantage over the circle and sector diagram because it is easier to place several bars within the space of a single graph than to include several circles in the corresponding space. The bars, furthermore, may have a common origin with reference to a percentage scale, whereas each circle must be more or less of an independent unit. The subdivided bar method is illustrated in chart 8.



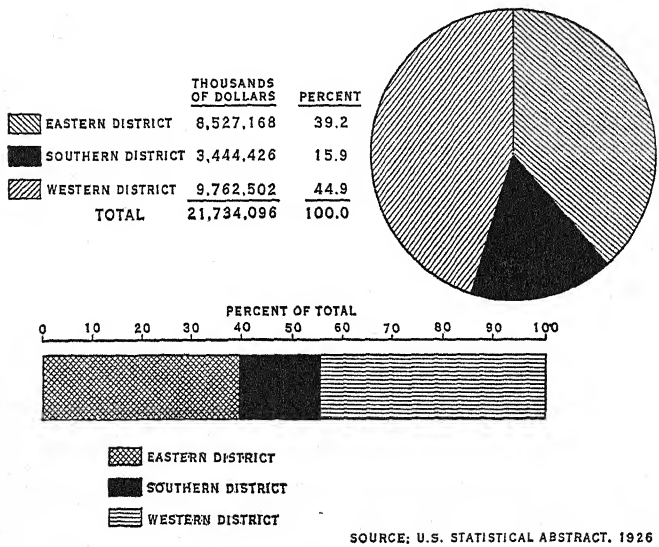
When a total is broken up into several components, it is generally more satisfactory to use the circle and sectors than the bar. In the bar illustration of chart 9 any two contiguous sections of the bar may be compared one with the other because they have one common boundary line, and the first and last

sections may be compared with the total for a similar reason; but comparison of the middle section with the total or the first with the last section is difficult without careful reference to the percentage scale. In the circle and sector diagram, on the other hand, there is one point common to every subdivision of the circle, namely, the center. All subdivisions focusing thus

CHART 9

RAILWAY SECURITIES OUTSTANDING

BY DISTRICTS, DEC. 31, 1925



SOURCE: U.S. STATISTICAL ABSTRACT, 1926

in a common point are readily compared, the one with the other and each with the total circle. The relative widths of the angles of the various sectors are judged with high accuracy as judgments of geometric measures go. The circle and sector diagram, therefore, possesses a distinct advantage over the subdivided bar where three or more subdivisions are involved.

This type of graph showing component parts has been used with great effectiveness in popular presentation of facts. Familiar examples are the dollar showing where the taxpayer's money goes, or the consumer's dollar showing the allocation to expense and profit of the selling price of beef or of some other similar commodity. This graph, in common with others, may become so complicated as to lose its effectiveness if an attempt is made to show too many components. It is probable that six or eight subdivisions represent about the maximum that can be shown satisfactorily by the average maker of graphs, though the Census Bureau sometimes makes as many as fourteen subdivisions.

Corresponding to the case shown above, page 81, where two subdivisions were shown for each of several totals, it is also possible to employ the graphic method for an analogous comparison of several components. Though the circle and sector diagram has the advantage over the bar diagram in the case of a single total so divided, there is possibly a more even choice between them when the attempt is made to show several such comparisons on the same graph. The two alternatives are shown in charts 10 and 11. The figures upon which these charts are based are given herewith:

UNITED STATES EXPORTS TO EACH CONTINENT, 1926

PERCENTAGE DISTRIBUTION BY STAGE OF MANUFACTURE

Continent	Crude Materials	Foodstuffs	Semi-Manufactures	Finished Manufactures
North America	16.5	19.7	14.0	49.7
South America	3.6	9.8	15.3	71.3
Europe	38.4	22.5	13.2	25.9
Asia and Oceania	22.7	7.0	16.0	54.3

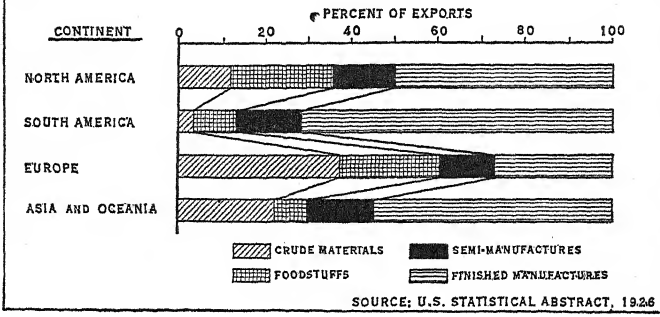
Source: *United States Statistical Abstract*, 1926.

Chart 10 has the advantage that the bars are close together and the guide lines between corresponding subdivisions of the

CHART 10

UNITED STATES EXPORTS TO EACH CONTINENT, 1926

PERCENTAGE DISTRIBUTION BY STAGE OF MANUFACTURE

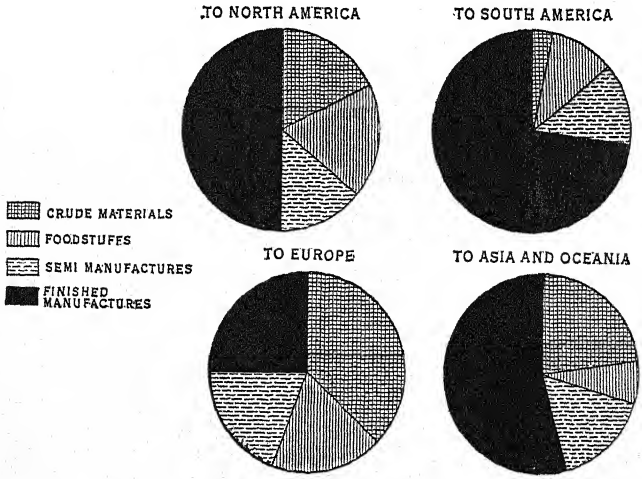


SOURCE: U.S. STATISTICAL ABSTRACT, 1926

CHART 11

UNITED STATES EXPORTS TO EACH CONTINENT, 1926

PERCENTAGE DISTRIBUTION BY STAGE OF MANUFACTURE



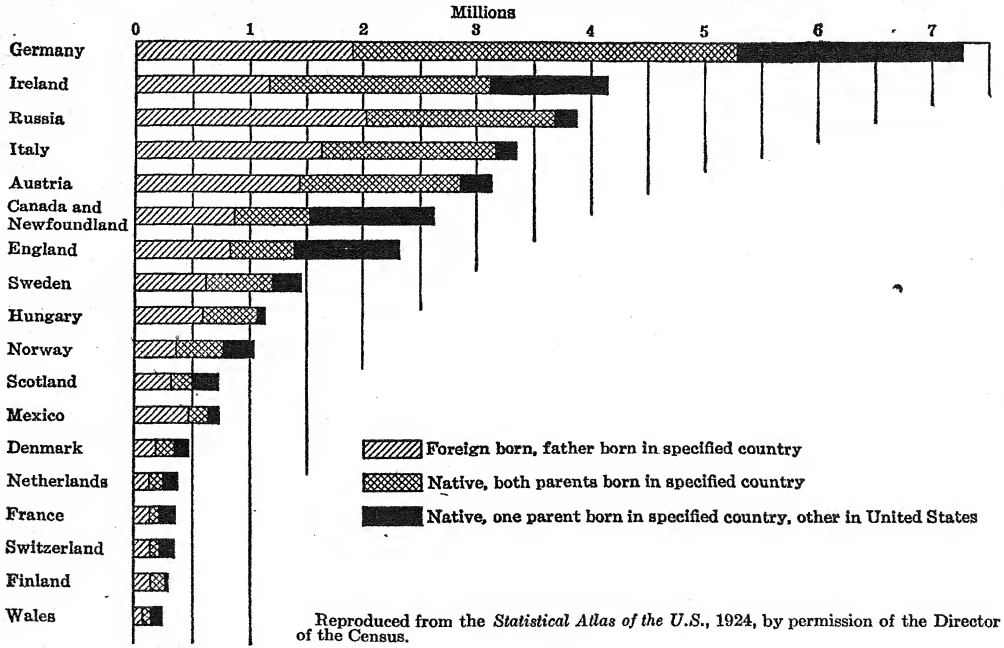
SOURCE: U.S. STAT. ABSTR., 1926

several bars assist in associating similar subdivisions and the bars all have the same lateral position with reference to the percentage scale at the top. It is possible to show the subdivisions of a large number of totals by means of the bars. The Census Bureau has no difficulty by this method in showing on a single page the bars for each of the forty-eight States. With the circles this association with a single scale is impossible; each circle must have its own scale, and a much smaller number of circles than bars can be shown on a single page. The subdivisions follow the same order in each circle; and despite the advantages of compactness and relationship to the scale possessed by the bars, it is doubtful if the relative magnitudes of the various parts to each other and to the total are as well shown by the bars as by the circles. So far as usage is concerned, the preference between these two methods has probably been in favor of the bars; but the United States Census Bureau uses both forms, with apparently little choice between them.

In presenting data such as those given in the last illustration, charts 10 and 11, it is frequently desired to show not only the varying components of the several totals but differences in absolute magnitudes as well. This again comes close to overreaching the possibilities of the graphic method, but at the same time can be done with a reasonable degree of success. Both the bar diagram and the circle and sector method may be used for this purpose. Charts 12 and 13, reproduced from the *United States Statistical Atlas*, 1924, illustrate the two methods, the former showing subdivision of the foreign white stock of our population into three classes and the latter the distribution of the total population for each census period, 1850 to 1920, into color and nativity classes. The bar graph drawn to a scale of actual populations permits comparison of the various totals one with the other, and the subdivisions of each bar can be compared approximately in terms of absolute numbers, but the pro-

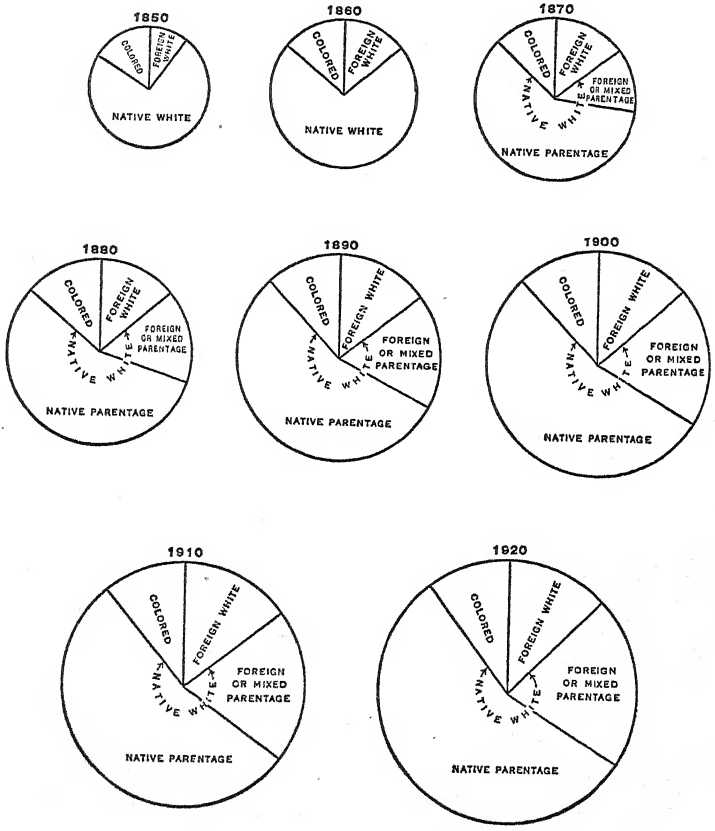
CHART 12

FOREIGN WHITE STOCK BY PRINCIPAL COUNTRIES OF ORIGIN: 1920



Reproduced from the *Statistical Atlas of the U.S.*, 1924, by permission of the Director of the Census.

CHART 13 TOTAL POPULATION AND DISTRIBUTION, BY CLASSES
1850–1920



Reproduced from the *Statistical Atlas of the U.S.*, 1924, by permission of the Director of the Census.

portion of each subdivision to its total can be estimated for only one bar at a time. It would be very difficult to say, for instance, whether the *proportion* of foreign-born to total foreign white stock were greater for Germany than for Ireland, or whether it were greater for England than for the Netherlands. In this respect the circle and sector graph possesses a distinct advantage over the bar graph, for the various circles are constructed of different sizes so that their areas are proportional to the various census populations, and yet the various sectors are directly comparable on a relative basis; that is, relative proportions are measured directly by the angles of the various sectors.

Note. In the matter of comparing the effectiveness of the subdivided bar with that of the circle and sector diagram, the results of several studies have been published recently. See (1) W. C. Eells: "Relative Merits of Circles and Bars for Representing Component Parts," *Journal, American Statistical Association*, June, 1926; (2) R. von Huhn: "A Discussion of Eells' Experiment" and F. E. Croxton: "Some Additional Data," both in *Journal, American Statistical Association*, March, 1927; (3) F. E. Croxton and R. E. Stryker: "Bar Charts *versus* Circle Diagrams," in *Journal, American Statistical Association*, December, 1927.

EXERCISES

I

Compare graphically (1) the numbers of men and women in the class; (2) the numbers belonging to each class in college; (3) the numbers from each department.

II

Draw a graph showing the numbers of representatives in Congress from each State. Data in *United States Statistical Abstract*, 1928, page 162. What would be the objection to representing the above data by a circle and sector diagram?

III

The *United States Statistical Abstract*, 1928, page 419, gives a table showing the number of vessels built in the United States by classes

and regions for each of several years. (1) Draw a graph showing the tonnage of the different classes built in any given year; (2) Draw a graph showing the tonnage built by regions for the same year. Graphs (1) and (2) may each be drawn in two different ways, representing two slightly different points of view or different interests in the data. What are the two ways and what is the difference in meaning? (3) Draw a graph which will make the comparison in (2) for each of two years. Are there alternative methods of doing (3)? If so, explain; if not, why not?

IV

Critical examination of magnitude and component parts charts. The *Statistical Atlas of the United States*, 1924, contains numerous illustrations of charts, some of complex character, that may be examined and criticized. It is also fairly easy to obtain illustrations of magnitude graphs, especially picture-representations, from newspapers and periodicals and these illustrations offer excellent opportunity for critical comment.

III

GRAPHIC REPRESENTATION OF FUNCTIONAL RELATIONSHIPS — CURVES

(*a*) INTRODUCTORY

The data of statistical curves. Statistical curves are used for the graphic representation of frequency distributions in the narrower meaning of the term and of time series. Logically any distribution of an aggregate or population into a specified set of categories or classes is a frequency distribution, and the categories may belong to any one of the four fundamental types of classification, quantitative, qualitative, geographical, or temporal. A population of human beings, for instance, may be classified by age or income class and a *quantitative* distribution results; or it may be distributed into nationality or sex or occupational classes — each a qualitative distribution; it may be classified by States or wards or sales districts; that is, geographically. The total week's attendance at the State fair may be classified by days of the week — a temporal distribution. Each illustration involves the distribution of a total or population into the subdivisions of a classification and the sum of all the class frequencies equals the whole population. Each then is properly called a frequency distribution. But usage of the term *frequency distribution* has been confined almost exclusively, in the literature of statistics, to the first case, namely, to the classification of frequencies into quantitative categories, and it is with this restricted meaning that the term is used in the following pages. Whether it be desirable to extend the use of the term to the broader meaning is a question that need be considered no further here. But about the desirability of having a term to apply specifically to distributions of frequency into

quantitative or magnitude categories there can be no question. Frequency distributions of magnitudes stand in a special class among statistical variates because they furnish statistical evidence upon an important type of scientific question, the question of the functional relationship between frequency and magnitude. Chaddock [1] gives an example of a frequency distribution of the weights of one thousand Freshmen which will illustrate the point:

Weight Class [a]	Number in class
90–	13
100–	28
110–	146
120–	245
130–	242
140–	160
150–	89
160–	46
170–	18
180–	9
190–	3
200–	1
	1000

[a] The figure designating each class is the initial value of the class; the intervals, therefore, should be read, 90 and under 100; 100 and under 110; etc.

The one thousand Freshmen are distributed into equal-sized weight classes, each a ten-pound interval, and the result permits a comparison of variation in frequency (numbers in each class) with variation in class size; thus there are thirteen students in the first class, and each step upward (by classes) shows an increase in frequency until the fourth class (120–) is reached; thereafter a decline. The farther a class is located from this position of maximum frequency the smaller the

[1] *Principles and Methods of Statistics*, 57.

number of frequencies included therein, and the decline is fairly uniform and regular in either direction. This relationship of frequency variation to magnitude variation is revealed only when the magnitude classes are all of the same size; if any class were a half or a third the size of others, its frequencies would be redistributed in some undeterminable way, and those for the odd-sized class would not then be comparable directly with the frequencies of full-sized classes.

The basic reason for distributing an aggregate or population in this fashion into a frequency distribution is to determine whether uniform and regular variation in frequency is associated with variation in class size; to determine, in formal language, whether frequency is a function of magnitude. The frequency distribution is, therefore, referred to as an expression of functional relationship. A frequency distribution is thought of ordinarily as a step in analysis preliminary to the calculation of averages, measures of dispersion, and the like; and this is correct. But the averages and dispersion constants are not likely to be very *constant* unless the values from which they have been calculated belong to a homogeneous group. And to say that they belong to a homogeneous group is but another way of saying that in the frequency distribution of these observations there will be some sort of regular or uniform variation in frequency associated with variations in magnitude, that the observations will tend to reflect a law of variation. It is true, of course, that frequency distributions are made up in many cases where this homogeneity or this tendency of the frequencies to follow any law of variation is conspicuously absent. The justification for them is a negative one, to show lack of homogeneity and the consequent instability of such series. Most economic data of the sort that is found to-day in records of governmental statistical bureaus or of business firms or associations are of the latter sort; they are characterized by a high degree of heterogeneity. This is no doubt one reason why the

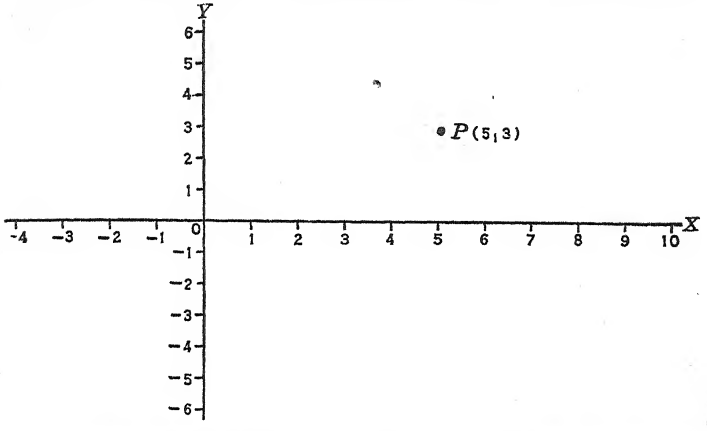
frequency distribution plays a lesser part in the analysis of current economic data than other forms of classification. But it must be emphasized that the frequency distribution is a fundamental method of analysis, and, even in its negative aspect, deserves an important place because of its capacity to reveal elements of heterogeneity. A frequency distribution of wages in a given plant or industry furnishes an excellent example of this fact. Not only may it reveal a lack of smoothness in the distribution of frequencies of wages, but it may show concentration points of groups of wages and thereby point to particular effects from heterogeneous elements. If average wages are calculated from such a group and then are used as the *typical* wages of the group, as is sometimes done, the heterogeneous character of the data suggests a high degree of unreliability of the average, and an unreliability in no way associated with the much-used probable error, or error of sampling.

As the frequency distribution points to a functional relationship between frequency and magnitude, so a time series points to a functional relationship between a statistical variate and time. The variate in the latter case, however, may be frequencies or it may be degrees of a measurable characteristic or it may be a magnitude derived from either of the two. Thus one time series expresses the dollar sales of a firm by months for the year — strictly, a distribution of frequencies in time, another, the population of a community by years — a time series of frequencies but not a temporal distribution. The temperature readings at a given station on specified dates illustrates the second general class above, while various averages, rates, or ratios illustrate the third class. The average temperature readings by months for a year may be obtained by averaging the daily highs and lows for each day of each month; per-capita sales or income, or production by days, months, or years are obtained by taking ratios of one aggregate to a related one for each time period.

Present knowledge of the functional relationships that exist in economic time series is not such as to enable one to list them with any assurance of finality. Certainly no general statement is possible that is applicable to all fields of scientific inquiry or even that is adequate for the social sciences as a whole. In economics (with possible extensions to some other social sciences) there have been attempts to set forth certain types of variation in time series corresponding to hypotheses as to the character of the causative forces in operation; but even these are looked upon as tentative formulations and may be subject to revision and change. The present work is not the place to consider the matter at length, and it suffices here to enumerate the various types of variation that are the subject of present study, for it is these that the graphic method is called upon to display. These fluctuations are usually classed as secular trend and cyclical, seasonal and irregular fluctuations. The problem of graphic representation of time series becomes the several problems of showing these different types of variation.

Graphic methods of representing functional relationships. Both frequency distributions and time series thus represent formulations of, or a search for, functional relationships. In the graphic representation of facts looking toward these ends, statistics borrows from mathematics as indeed it does in many other respects. Two lines are drawn in a plane, intersecting at right angles, as in the diagram given herewith, and scale figures are drawn on each line indicating distances in each given direction from the point of intersection or origin, 0. The vertical line is called the y-axis, or axis of ordinates, and the horizontal line the x-axis or axis of abscissæ. They are referred to jointly as coördinate axes; and distances in the plane measured along the y-axis — that is, distances in a vertical direction from the origin — are called ordinates, while distances measured horizontally from the origin along the x-axis are called abscissæ. It is possible to describe exactly the location of any point in the

COÖRDINATE AXES AND THE REPRESENTATION OF A POINT IN A PLANE



plane by giving its position with reference to the coördinate axes. The point P is five units distance to the right of the origin and three units distance above it; its abscissa is 5, its ordinate 3. By convention, the location of any point is indicated by naming first its abscissa and then its ordinate; that is, P (5, 3) refers to the point shown in the diagram, whereas P (3, 5) would refer to a second point with abscissa 3 and ordinate 5. This principle of representing paired variables by points in a plane is basic to the graphic representation of the relationship of two statistical variates.

Exact functional relationships between variables are expressed in the form of equations, an equation between two variables for instance expressing the fact that, and the way in which, one variable is uniquely determined by the other. Some of the simpler forms of equations are:

$$y = ax$$

$$y = a + bx + cx^2$$

$$y = \frac{a}{x}$$

$$y = ab^x$$

In each of these cases, x is called the independent variable, signifying that it may vary independently; while y, the dependent variable, varies only in response to variation in x and in a predetermined way — in accordance with the formula expressing the relationship between them. In the first case above, for instance, y varies according to a fixed multiple, a, of x. Thus if a has a numerical value of eight, a variation of five units in x will produce a variation of eight times five, or forty, units in y.

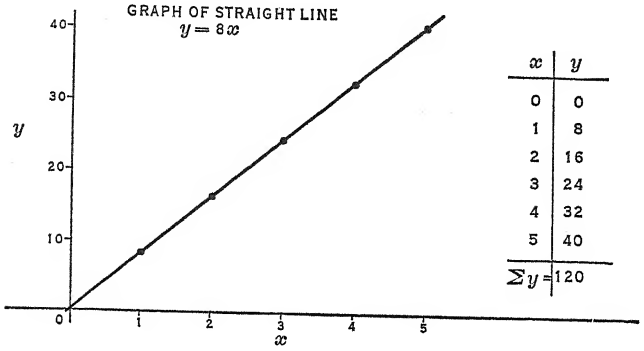
No series of economic data is found, of course, in which the relationship between two observed variables is of this exact and unchanging nature. Attention has already been called [1] to the fact that the data of economics, and indeed of the social sciences generally, are the resultant of a complex set of causes, interacting one with the other. But it must not be concluded from this fact that the equations of mathematics, expressing exact functional relationships, are of no use to the economist, for cases frequently arise in which the influences of one causal factor, or of one set of causal factors, sufficiently predominates in a given series to enable the series to be represented approximately by some kind of equation.

It will be worth while, therefore, before considering the graphic representation of statistical variates (frequency distributions or time series) to look briefly at the graphic representation of those exact functional relationships represented by some of the equations given above. Suppose it be desired to represent on a diagram the functional relationship between x and y expressed by the formula $y = ax$. Let a be given a numerical value of 8; then when x varies one unit, y varies eight units; if x varies five units, y correspondingly varies forty units; and so on. The table of values in chart 14 shows values of y for integral values of x varying from 0 to 5. The relationship of x and y may now be shown graphically. A pair of coördinate axes is drawn and on the x —, or horizontal, axis a scale is indicated to

[1] See pages 1 to 6.

include all the values of x for which the curve is to be shown; a corresponding scale on the vertical axis for y. The points given in the table of values are then plotted, that is (0, 0), (1, 8),

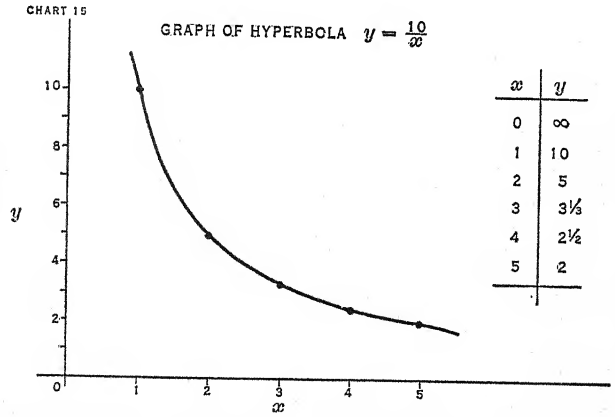
CHART 14



x	y
0	0
1	8
2	16
3	24
4	32
5	40
$\Sigma y =$	120

..... (5, 40). A line is then drawn connecting the points. The characteristic of the linear function, $y = 8x$, clearly evident

CHART 15

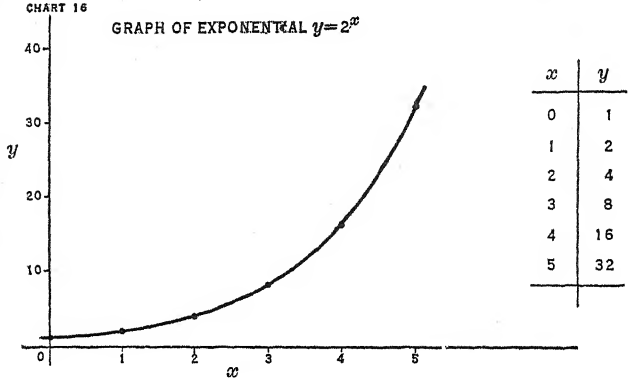


x	y
0	∞
1	10
2	5
3	3⅓
4	2½
5	2

from the graph, is the nature of the change in y as x changes. Starting from any position in the curve, as it moves to the right

it moves upward at the rate of eight units of vertical distance for one unit of horizontal distance; the amount of change in the

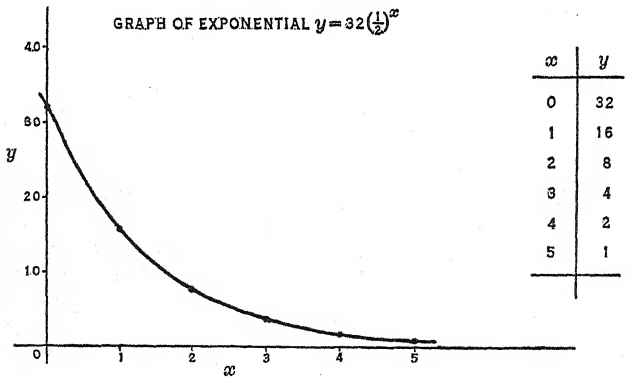
CHART 16



x	y
0	1
1	2
2	4
3	8
4	16
5	32

curve per unit of time, assuming that x represents time, is constant, or, otherwise stated, the curve has a constant slope.

CHART 17



x	y
0	32
1	16
2	8
3	4
4	2
5	1

For the curves $y = 10/x$ and $y = 32\ (\frac{1}{2})^x$ (charts 15 and 17) the value of the function declines as x increases, but they differ

from the straight line in that the amount of change for unit change in x varies for different positions on the curve. In the early portions of the curve the decline is rapid, in later portions much slower. The graph of $y = 2^x$ (chart 16) shows an upward-sloping curve with the slope increasing as the curve moves to the right.

(1) The characteristic of the graphic representation of these several functions that attracts immediate attention is the ease and accuracy with which their slopes can be compared at different points. The eye is able to judge with a high degree of accuracy the question whether a curve is increasing or decreasing more rapidly at one point than at another. This ability to judge the slope of the curve is closely associated with the ability previously emphasized [1] to judge the size of two angles, for the judgment of slope is, in final analysis, a judgment of angular measurement. For if, at a given point on the curve where the slope is to be determined, a ruler be laid tangent to the curve, that is, it is to touch the curve but not cross it at that point, and if the ruler be extended to cross the horizontal axis,[2] the slope of the curve is then determined by the size of the angle which the ruler makes with the horizontal axis — the larger the angle the greater the slope, and *vice versa*. If the curve moves upward to the right, the slope is considered positive; if upward to the left, the slope is negative.

(2) Another characteristic of each of these curves which is made clear by the graph is the continuous character of the variation for the part of the curve shown on the graph. There are no breaks in any of the curves. For the purpose of plotting, values of y were calculated only for integral values of x varying from zero to five, and the curves were then drawn freehand between these plotted points. Values of y could equally well have been calculated for many fractional values of x, but it would not

[1] See pages 65 and 66.

[2] The only case in which this extended line will not cross the horizontal axis at some point is where the tangent to the curve is parallel to this axis.

have resulted in a graph that showed to the eye any significant variations from the graph as drawn. To say that a function, y, is continuous from $x = 0$ to $x = 5$ is to say, in simple non-technical language, that a real value exists for y corresponding to any value whatsoever of x between these limits, whether it be integral or fractional. If it were desired to show discontinuous variation in the relationship of two variables, then there should be plotted on the graph only the points at which values of y were found associated with values of x, and the result would be a series of points or a broken line. If, for instance, a function existed such that the values 0, 8, 16, 24, 32, 40 of y were found for values of $x = 0, 1, 2, 3, 4, 5$, but no values of y for any other value of x, such as $\frac{1}{2}$, $2\frac{1}{4}$, etc., then y would be a discontinuous function of x, and the graph of this function would be represented by the points plotted on chart 14.

The graphs of these equations possess a third characteristic which is important in understanding one feature of the curves and which often has significant applications when graphs constructed on similar principles are used to represent statistical aggregates or frequencies. If a rectangle be drawn in the plane with sides the length of one unit on the x- and the y-scales respectively, the area of this rectangle may be called a unit of area. If now the values of the dependent variable, y, be summed (integrated) for values of x between given limits, the total sum of the y's will equal the number of units of area in the space bounded by the base line (x-axis), the curve and the two ordinates erected at the limiting values of x between which the summation was made. For instance, in chart 14, the number of units of area under the curve from $x = 0$ to $x = 5$ equals 100, as can be ascertained easily by inspection, since the curve is a straight line running diagonally through a rectangular area, five units by forty units. Of more importance in this case will be a comparison of area units with the values of y given in the table for values of x from zero to five; but some care is necessary to

make this comparison properly. In this case the integral values for x shown in the table should be considered as class values at unit distance apart on the x-scale; that is, $x = 0$ represents a distance from $x = -\frac{1}{2}$ to $x = +\frac{1}{2}$; $x = 5$ represents the distance from $x = 4\frac{1}{2}$ to $x = 5\frac{1}{2}$. Making the necessary calculations for estimating the total area under the curve, therefore, from the limits $x = -\frac{1}{2}$ to $x = +5\frac{1}{2}$, for comparisons with the sum of the y's in the table on chart 14 and noting that the areas below the x-axis are considered as negative, it is found that the total area within the limits given is 120, agreeing with the sum of the y's in the table. This property of the curve, that the total area under the curve equals the sum of the values of the dependent variable, has its most important applications in the statistical graphs of frequency distributions, but is sometimes of use also in distributions of frequencies in time.

N.B.

EXERCISES

I

Construct a graph of the equation $y = 10 + .5x$ for values of x from zero to eight. (*Hint:* Make a table of values of y for integral values of x from zero to eight; plot these values on the graph and draw curve of equation freehand between plotted points.)

II

Construct a graph of the equation $y = 5 + 2x + 2x^2$; a graph of the equation $y = 5 + 2x - 2x^2$, each for values of x from zero to eight.

III

Construct a graph of the equation $y = 6 + x + 2x^2 + x^3$ for values of x from zero to eight.

IV

Construct a graph of the equation $y = \frac{a}{x}$ for values of x from one to six, when $a = 60$; when $a = -60$.

V

Construct a graph of the equation $y = 3^x$ from $x = 0$ to $x = 5$. (*Hint:* $3^0 = 1$.)

(*b*) GRAPHS OF FREQUENCY DISTRIBUTIONS

Graphs of simple frequency distributions, continuous series. Frequency distributions show how an aggregate of frequencies are distributed along a scale of magnitude. The magnitude scale is first divided into a number of equal intervals or classes and the numbers of frequencies that fall into each class are then shown. An example is the numbers of wage-earners of a given group receiving weekly wages of five and less than ten dollars, of ten and less than fifteen, etc. Sometimes it is important to distinguish between continuous and discontinuous variation. When variation is continuous, individual cases may be expected to have any value on the magnitude scale between the highest and lowest limits recorded; if the variation is discontinuous, or if the series in question is discrete, as it is sometimes called, there will be gaps on the magnitude scale within which no frequencies will be found. In a frequency distribution of the ages of a population, for example, though the population may be classified into five-year age groupings, it is recognized that any individual in the "twenty and less than twenty-five" group may have at a given instant any conceivable age between these limits. In classifying the population into these groups, therefore, much of the detail with reference to their exact ages is lost, but the fact is not lost sight of that age is a continuous variable. If again one were engaged in the age-old experiment of tossing ten coins repeatedly and in recording as successes at each toss the number of coins falling heads up, one would obtain from this experiment a discontinuous frequency distribution. The number of successes in any toss must always be integral; each time it must be one of the values in the series, 0, 1, 2,..... 10. To get any fractional number of successes is impossible, hence the gaps in this distribution of values involve all except integral values on the magnitude scale. At times it is appropriate to consider a variable as continuous when in reality it is discontinuous, but its discrete character arises out of customary

methods of quoting variables. For instance, there are few instances where any importance attaches to the consideration of a frequency distribution of wages as discrete, although the wages in question may never involve other than even dollars and can never involve fractional parts of a cent, since in ordinary payments the cent is the smallest subdivision of our monetary system. Interest rates generally are discrete variables, and it is often desirable that their discontinuous character be emphasized.

It is desired, then, to have a graphic method of representing a continuous distribution of frequencies. Take for example the following hypothetical data from the payroll of an industrial establishment:

WEEKLY EARNINGS OF WAGE-EARNERS

PAYROLL PERIOD ENDING OCTOBER 20, 1928

Wage class[1]	Number of workers
$10–	65
15–	143
20–	118
25–	87
30–	42
35–	5
Total	460

[1] The figures given in the wage-class column refer to the initial values of the class. $10–, for instance, means $10 and less than $15. This method of designation, where initial values of each class are given in the table, will be followed rigidly hereafter. When a class figure is given but not followed by the dash, it will refer to the mid-value of the class. Thus the above figures, if they were not followed by the dashes, would refer to class limits as follows:

$7.50 to $12.50
12.50 to 17.50 etc.

The methods developed for the graphic representation of the functional relationship between two variables may be utilized in portraying the facts in this frequency distribution. The total space to be occupied by the chart having been decided

upon, the distribution of this space as between title and actual graph is decided on the basis of considerations repeatedly urged in the previous pages. The first step in the utilization of the space allotted to the graph is to construct vertical and horizontal scales, the vertical to represent frequencies, or numbers of workers, and the horizontal, magnitudes, or wages. The problem of using properly the space to be devoted to the graph is solved by the proper arrangement of these scales. Noting that the maximum and minimum limits of the wage classes given in the distribution are \$40, the upper limit of the largest class, and \$10, the lower limit of the smallest class, the horizontal scale is so arranged that the major portion of the horizontal space will be taken up by the scale positions from 10 to 40, allowance being made of course for proper margins on either side. It is not necessary in general that the zero position on the magnitude scale be shown on a frequency diagram, although in this case it can easily be included. The vertical scale, for frequencies, should be set so that the highest class frequency given in the distribution, in this case 143, will extend nearly to the upper limit of the space allotted to the graph. This scale must *always* begin at zero.

The scales having been set, the graph representing the facts of the frequency distribution may be drawn in one of three ways, the resulting figures being called respectively (1) the *histogram*, (2) the *frequency polygon*, and (3) the *frequency curve*. To construct the histogram, plot a point in the middle of each class interval — that is, with abscissa equal to the mid-value of the class — and at a height, or with an ordinate, equal to the number of frequencies in the class. Draw lines through each of these plotted points parallel to the x-axis and extending in each direction from the point to the limit of the class interval. Connect these lines by ordinates drawn perpendicularly from one to another and to the base line, the ordinate lines all lying on class boundaries. The result is an area completely bounded

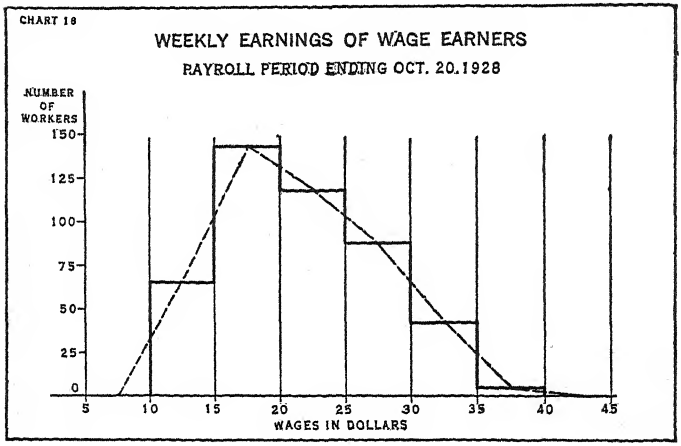
by the ordinates, the base line and the parallels at the top of each class. Sometimes each ordinate is extended to the base line so that the figure is composed of a set of rectangles one over each class interval on the magnitude scale. Three characteristics stand out as features of this method of representing the frequency distribution. In the first place, recalling the significance of area-representation in this kind of diagram, the unit of area being measured by a rectangle one class interval in width and one frequency unit in height, the total area of the diagram included within each class interval is exactly equal to the number of frequencies in that class of the distribution. The second characteristic follows directly from this: the area of the entire figure is equal to the total number of frequencies in the distribution. In the third place, the heights of the various rectangles give an immediate and accurate impression to the eye of the entire distribution of frequencies among the several classes of magnitude; the maximum class is prominently shown and also the fairly regular and uniform decline in frequencies as the classes progress to the right of this position of maximum frequency. The procedure is illustrated by the unbroken line in chart 18.

The construction of the frequency polygon[1] differs from that of the histogram only in that, after the points are plotted at the mid-positions of each class interval as before, they are now connected by straight lines drawn from each point to the next, the diagram being completed by drawing a straight line from the point at the top of the first class interval directly to the base line at the mid-point of the previous interval,[2] and similarly

[1] There is an inconsistency in confining the term *frequency polygon* to this diagram, for the histogram is a frequency polygon as well, though of a special sort. It is sometimes called a rectangular frequency polygon, sometimes a block diagram or a column diagram. The usage here adopted follows Yule, who refers to Pearson. See Yule: *Introduction to the Theory of Statistics*, chapter 6.

[2] The frequency polygon may in one instance here do violence to the facts. Suppose that the first class were "0 and under 5" and that negative frequencies

from the point in the last class interval to the base line at the mid-point of the next succeeding class interval. This procedure produces an area-representation of the frequency distribution that differs slightly from the preceding one. In the several class-intervals a triangular strip of area is cut off here and added there; but for each area cut off, an exactly equal area is

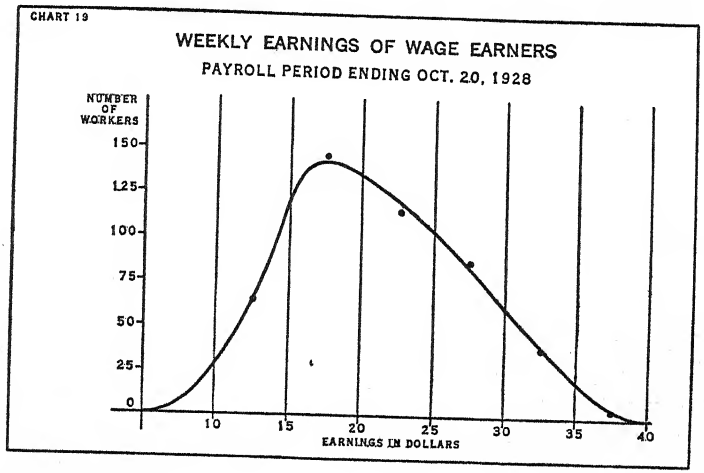


added, although not in the same class. In the diagram, the broken line of chart 18, the amount added in the interval five to ten, in which no frequencies appeared in the actual distribution, has been taken away from the upper left-hand corner of the rectangle in the class, ten to fifteen. These two areas, the one added and the other subtracted, are identically equal; and similarly for the remaining classes of the diagram. It follows that the frequency polygon has preserved one characteristic of

were not found. Drawn as described, the polygon would extend from zero frequency at $-2\frac{1}{2}$ on the magnitude scale to a height at the center of class one representing the number of frequencies in this class. But this representation indicates frequencies at negative values on the x-scale. This difficulty can be avoided only by bringing the polygon to the base line at zero magnitude and so adjusting its early course as to preserve equality of areas and frequencies.

the histogram, namely, the equality of total area and total frequencies; but has destroyed another, for the class areas of the polygon do now agree with the class frequencies of the original data. The frequency polygon, in other words, has brought about a readjustment or rearrangement of the frequencies. The significance of this change will be considered presently. The uniformity of the decline of frequencies on either side of the point of maximum frequency is greatly accentuated by the second figure.

To illustrate the construction of the frequency curve, a second chart, number 19, has been drawn. Up to the point of



setting the scales and drawing the points representing the frequencies in the several classes, this chart is identical with chart 18. The curve in this case has been drawn freehand, the purpose being, not to make it pass exactly through each of the plotted points, but as near to them as possible consistently with the requirement that there shall be no sudden breaks or changes in direction to the curve, to present a picture of regular and

uniform variation in magnitude. A fundamental characteristic of this curve, which it is difficult to realize except approximately when the drawing is done by freehand method, is that the area under the curve shall represent the total number of frequencies in the entire distribution.

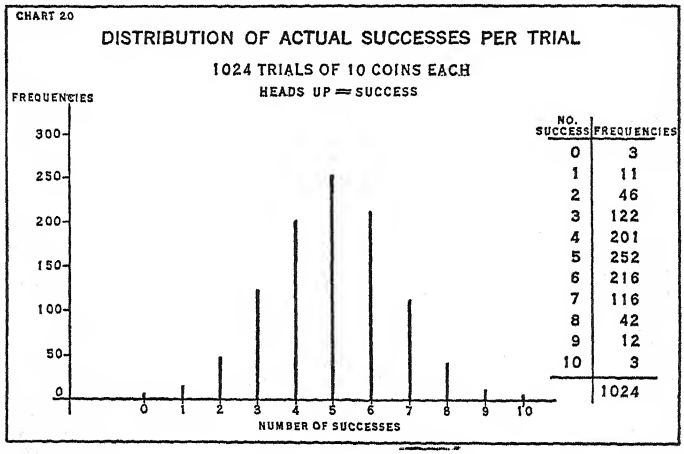
A comparison of the uses of the three types of representation can now be made. The histogram is found to be a graphic representation of the actual facts of a given frequency distribution; it shows the distribution of the total number of frequencies in the several classes just as they have occurred in this particular instance. The frequency polygon is something more than, or at least something different from, a representation of these particular facts. It looks to these facts as in themselves a representation of a broader group of facts, a sample of a larger group or of a more general situation facing the particular group of workers. It may be stated thus for the earnings of this group of workers at this payroll period: they may not constitute the total class of workers in this period, but are taken as representative of the entire class; or the purpose may be to discover, for this particular group, what constitutes the general or typical distribution of earnings among them, and this payroll period is one bit of evidence, but may disagree in some details with the experience which is typical for the group when a longer period of time is considered. Herein lies the reason for smoothing the histogram, and the frequency polygon is the first step in that process. Its purpose is to discover what, from a particular instance, can be learned about a more general situation, and it is assumed that the irregularities of the particular instance are due to the broad categories of the classification or to particular and non-typical occurrences which crop up at all times and show themselves more prominently among a small number than would be the case with a larger group or in a longer period of time. If, then, the particular frequency distribution cannot be looked upon as a sample of a larger group or of a longer period of time,

there is no justification for smoothing — and none therefore for the frequency polygon and frequency curve. These two graphs represent generalizations from a particular set of facts, the former a first step in the process, or a half-way generalization, the latter a complete generalization. The frequency curve given in chart 19 was drawn freehand, but in a more thoroughgoing solution of the problem of generalization from a given instance would be drawn differently. If the frequency distribution is constructed from homogeneous data, it may be generalized in any one of several different ways; there are various laws of variation expressing frequency as a function of magnitude. The worker in a given field of science, the economist for example, must decide from a study of his data what law of variation it will follow, and the mathematical equation expressing this law may then be *fitted* to the actual observations, and if a good *fit* is obtained there is evidence to support the assumption that the given observations do represent the selected law.[1]

Simple frequency graphs, discrete series. When the data of a frequency distribution are discontinuous and it is desired to indicate their discontinuous character on a graph, the line through the tops of ordinates, as drawn in the histogram, polygon, and curve, is not appropriate to the representation, for this line in each of the three cases has completely enclosed an area on the base line of the diagram and is interpreted in terms of continuous variation. This may be illustrated by reference to the frequency curve. Although the actual frequency distribution shows frequencies only for certain *classes* of magnitude, for example, page 103, 143 workers each earning at least $15 per week and not as much as $20; yet the frequency curve permits an estimate of the numbers who would be found earning $15 and

[1] The study of these laws of variation and the description of methods of fitting transcend the scope of this book. It is here only necessary to recognize that whenever a smooth curve is fitted to a frequency distribution by freehand or otherwise, the curve carries implications that a law of variation exists and that this law is indicated by the given set of observations.

less than $16; or the numbers earning $16 and less than $16.50, etc., etc. If, however, a frequency distribution of interest rates is to be shown graphically, and if the purpose is to indicate the fact that interest rates are quoted in no smaller units than quarter per cents, a continuous curve will misrepresent the facts to be shown. The discontinuous character of the data may be represented by erecting ordinates at each point on the x-scale where cases are found, each ordinate being of a length determined by the number of frequencies of a given value.



In chart 20, the discontinuous character of the results of a coin-tossing experiment is indicated by a graph of this kind. Frequencies occur only at integral values on the magnitude scale, and an ordinate has been drawn for each integral value from zero to ten inclusive, the ordinate being of a length in each case equal to the number of tosses that resulted in the given number of successes, the total number of trials or tosses being 1024. Sometimes, instead of thin lines representing these ordinates, bars of greater width are erected for the purpose of

bringing the frequency representation into greater prominence. The conception of discontinuous variation is still preserved.

This method of graphic representation of a discontinuous frequency distribution is resorted to only when it is desired to emphasize the discrete character of the series. Instances frequently arise in which the variable may actually be discontinuous and yet the purpose of the graphic portrayal is properly served by the usual histogram, polygon, or curve. This is especially true when the discontinuity is of a formal character, arising from the customary units in which the variable is quoted.

Cumulative frequency distributions and graphs. Some of the peculiarities of frequency distributions are more easily seen when the distributions are given in *cumulative* form, and there are appropriate graphic methods to use with the cumulative distribution. Consider again the data on page 103, reproduced herewith.

WEEKLY EARNINGS OF WAGE EARNERS

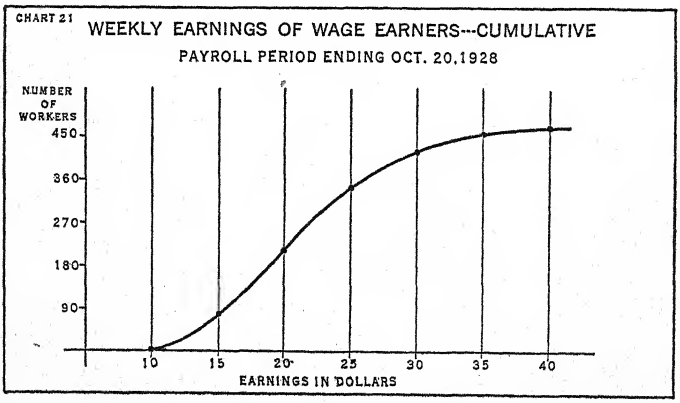
PAYROLL PERIOD ENDING OCTOBER 20, 1928

Wage Class	Number of Workers	Cumulative Number of Workers	Cumulative Percentage of Workers
10–	65	65	14.13
15–	143	208	45.22
20–	118	326	70.87
25–	87	413	89.78
30–	42	455	98.91
35–	5	460	100.00
Total	460	460	100.00

The third column of the table, showing cumulative numbers of workers, is obtained by summation of the successive items in column two. Thus there are 65 workers with wages less than

$15; 208 workers, that is, 65 plus 143, with wages less than $20, etc. The last column, giving these figures in percentages of the total number of workers, is especially valuable in showing how the workers are distributed with regard to earnings. Fourteen per cent receive less than $15; 45 per cent, less than $20; 90 per cent (89.78), less than $30. It is equally possible, and is in some cases desirable, to cumulate in the opposite direction; that is, from the highest to the lowest class: thus, 5 workers receiving $35 or over, 47 receiving $30 or over, and so on. And, in percentages, though not shown in the table, the results are 1 per cent receiving $35 or over; 55 per cent receiving $20 or better, etc.

Cumulative frequencies such as these may be plotted upon a diagram as follows: Construct the two scales as before, the magnitude scale on the horizontal to show all magnitudes from lowest to highest; the frequency scale on the vertical beginning as before with zero and extending to such a point that the *total* number of frequencies in the distribution can be plotted near the top of the graph. The method is illustrated in chart 21. The essential difference between this graph and graphs of simple frequency distributions lies in plotting the points represent-



ing frequencies for the various classes. Instead of plotting the point representing the cumulative frequencies of any given class at the mid-position of the class on the x-scale, it is plotted at the extreme limit of the class in the direction in which the cumulation has taken place. The reason for this is easily understood when the cumulative character of the frequencies is considered. Beginning at any point on the wage scale prior to \$10 and moving to the right, it is noted that no frequencies are encountered before the \$10 position is reached. Nothing is known, of course, about the distribution of the 65 frequencies in the \$10 class; but by the time this class has been passed and the point, \$15, reached on the magnitude scale, it is known that 65 frequencies have been encountered. These facts are represented graphically by plotting a point, zero frequencies, on the \$10 position of the magnitude scale and a second point, 65 frequencies, on the \$15 position of the x-scale; and correspondingly for the remaining classes. When the \$35 class has been traversed and the position \$40 reached, a total of 460 frequencies have been encountered, and the ordinate, 460, is therefore plotted on abscissa 40. These points are then connected by a smooth line; it is possible to draw a straight line between each two points, but in practice the other method is followed. The characteristic features of this graph may now be noted. First, while geometrically the area under the curve may have a significance corresponding to that of the simple frequency curves dealt with previously, this area has no significance in terms of the economic data represented; and this statement holds for all the cumulative frequency data in economics. Secondly, the curve permits of interpolation between the plotted points; for instance, to estimate the number of workers receiving earnings of \$27 or less, draw an ordinate through the position 27 on the x-scale and extend it to intersect the curve. From the y-scale read the value of this ordinate at the point of intersection and this is the required estimate.

If a percentage scale of workers be erected alongside the scale of numbers of workers or at the right of the diagram so that the 100 per cent ordinate is equal to the 460 workers ordinate, then it is easily possible to extend this method of interpolation to answer such questions as the following: below what earnings figure are the lowest one quarter, or one half, or two thirds of the workers found? The slope of the curve is important also, for it gives an accurate indication of the relative concentration of frequencies at different points along the scale of magnitude; where the slope is steepest — that is, rising most rapidly — the frequencies are most numerous, and where it rises slowly or flattens out, or where the slope approaches a zero value, the frequencies are few in number.

By modifying the form in which frequency data are arranged, a special form of cumulative frequency curve, known as the Lorenz curve, may be used effectively. It was first used by Dr. M. O. Lorenz to show distribution of incomes. Instead of showing only the frequency of income recipients in the several income classes, income data may be arranged to show in addition the total amounts of income received by the persons in each class, and these data may finally be expressed in terms of the percentages of the total group of persons receiving various percentages of the total income. The following data (Table 11, page 115), taken from *Income in the United States*, I,[1] 136–37, illustrates.

Chart 22 is constructed to show the facts in the last two columns of this table on distribution of incomes. Each scale is a percentage scale — of incomes on the horizontal and of persons receiving income on the vertical. The excellence of this type of graph lies in the ease and accuracy with which the curve of incomes indicates inequality in the distribution of incomes. If incomes in a community were distributed with complete equality, 10 per cent of the people would possess 10 per cent of

[1] Publication of the National Bureau of Economic Research.

TABLE 11.

DISTRIBUTION OF INCOMES IN THE UNITED STATES, 1918*

INCOME CLASS	SIMPLE DISTRIBUTION		CUMULATIVE DISTRIBUTION Percentage of Total Under Class Above.	
	Number of Persons	Amount of Income —000—	Percentage of Persons	Percentage of Income
Under zero	200,000	—$125,000	.5324	—.22
0—	1,827,554	685,288	5.3969	.96
500—	12,530,670	9,818,679	38.7506	17.90
1,000—	12,498,120	15,295,791	72.0176	44.30
1,500—	5,222,067	8,917,648	85.9175	59.69
2,000—	3,065,024	7,314,413	94.0759	72.31
3,000—	1,383,167	5,174,091	97.7576	81.24
5,000—	587,824	3,937,183	99.3222	88.03
10,000—	192,062	2,808,290	99.8334	92.88
25,000—	41,119	1,398,786	99.9428	95.29
50,000—	14,011	951,530	99.9801	96.93
100,000—	4,945	671,566	99.9933	98.09
200,000—	1,976	570,019	99.9986	99.07
500,000—	369	220,120	99.9996	99.45
1,000,000—	152	316,319	100.0000	100.00
Total	37,569,060	57,954,722		

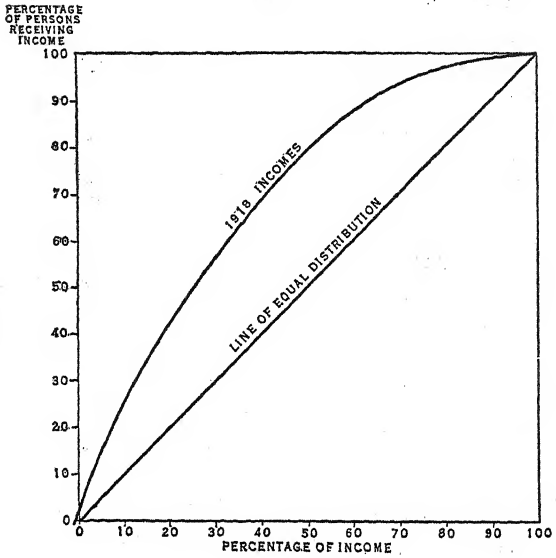
*Excluding 2,500,000 soldiers, sailors and marines.

the income, 50 per cent of the people 50 per cent of the income, and so on all along the scale. The divergence of the income curve, therefore, from the diagonal of the diagram indicates the extent of *inequality* in income distribution. The smooth curve drawn through the plotted points permits estimates of this inequality at any position; thus, 80 per cent of the group are estimated to receive 52 per cent of the income. The cumulation of the data in the table could have been made from the other end of the scale equally well.

Comparisons of frequency distributions, simple or cumulative. The simple and the cumulative frequency graphs have each in turn brought into prominence particular features of the distributions which they represent; for the simple frequency

CHART 22

DISTRIBUTION OF INCOMES IN THE UNITED STATES
1918



graph, the position of the maximum frequency and the decline in frequencies on either side of this maximum position; for the cumulative graph the positions above or below which any number or proportion of the frequencies fall and also the relative concentration of frequencies at different magnitudes as indicated by steepness in the slope of the curve. Sometimes it is desired to see how these characteristics compare in two related frequency distributions, such as wages in a given firm at two periods of time or the wages of two different groups at the same time. These comparisons may be made effectively by plotting the two curves on the same diagram, illustrated for simple frequency curves by chart 23 and for cumulative curves by chart 24. The following hypothetical data are used:

TABLE 12.

WEEKLY EARNINGS OF WAGE EARNERS

PAYROLL PERIODS FOR OCTOBER 1927 AND OCTOBER 1928

WAGE CLASS	PAYROLL OCTOBER 1928			PAYROLL OCTOBER 1927		
	Number of Workers	Percentage of Workers	Cumulative Percentage of Workers	Number of Workers	Percentage of Workers	Cumulative Percentage of Workers
$10—	65	14.13	14.13	72	16.00	16.00
15—	143	31.09	45.22	165	36.67	52.67
20—	118	25.65	70.87	112	24.89	77.56
25—	87	18.91	89.78	70	15.56	93.11
30—	42	9.13	98.91	31	6.89	100.00
35—	5	1.09	100.00	0	0	100.00
Total	460	100.00	100.00	450	100.00	100.00

Where the two distributions refer to materials of the same general character and where they contain about the same number of frequencies, they may be plotted on a diagram with the vertical scale expressed in *numbers* of frequencies without undue danger that the comparison will be invalidated. If, however, the numbers of frequencies in the two distributions differ to any considerable extent, it is important to express the frequencies in relative terms — that is, the class frequencies as percentages of the total — and to express the vertical scale in percentages. This has been done in charts 23 and 24. The graphs on chart 23 are drawn as frequency polygons and show clearly the important differences between the two distributions — the greater proportion of workers of the earlier period in the $10 and $15 classes and the smaller proportions in the higher wage classes. The cumulative graphs, because of the greater unfamiliarity of most persons with them, are somewhat more

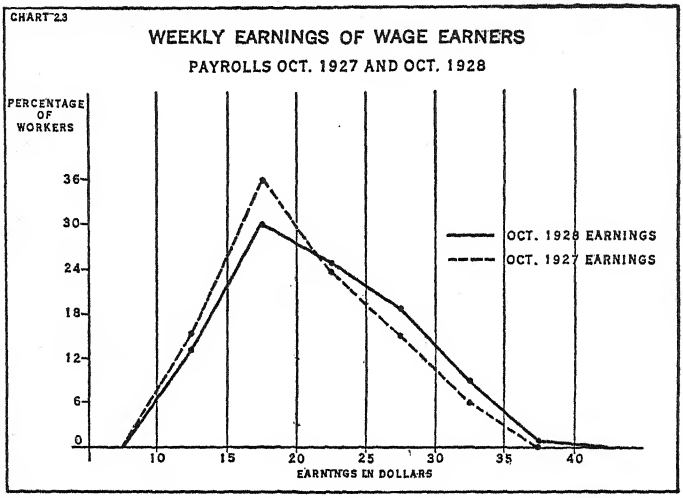


CHART 23
WEEKLY EARNINGS OF WAGE EARNERS
PAYROLLS OCT. 1927 AND OCT. 1928
PERCENTAGE OF WORKERS
36
30
24
18
12
6
0
10
15
20
25
30
35
40
EARNINGS IN DOLLARS
OCT. 1928 EARNINGS
OCT. 1927 EARNINGS

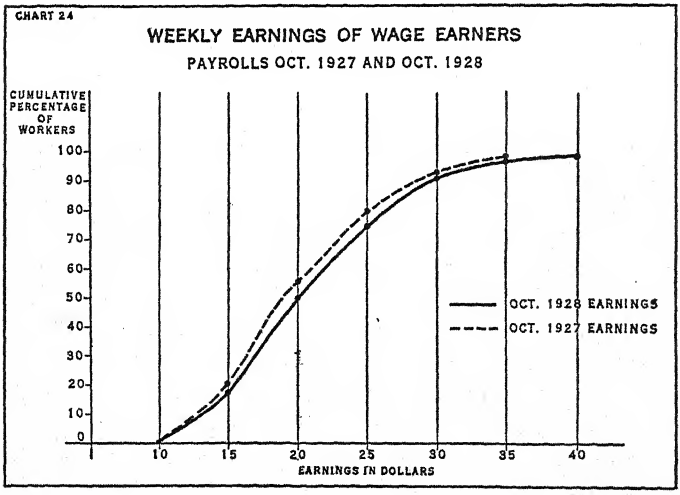


CHART 24
WEEKLY EARNINGS OF WAGE EARNERS
PAYROLLS OCT. 1927 AND OCT. 1928
CUMULATIVE PERCENTAGE OF WORKERS
100
90
80
70
60
50
40
30
20
10
0
10
15
20
25
30
35
40
EARNINGS IN DOLLARS
OCT. 1928 EARNINGS
OCT. 1927 EARNINGS

difficult to interpret. The relative positions of the two curves furnish the key to their interpretation. For a given ordinate, say $20, it is noted that the 1927 curve shows 52 per cent of the workers, whereas the 1928 curve shows but 45 per cent; that is, 52 and 45 per cent respectively earning under $20. The fact that the 1928 curve lies below or to the right of the 1927 curve indicates that throughout the distribution the 1928 earnings were higher than those of the earlier period — fewer people earning wages under a given amount; the earnings of a given proportion of the total being always higher.

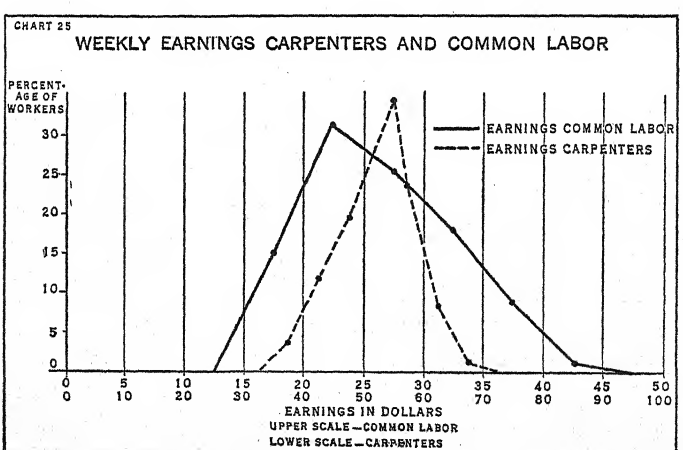
There is one other kind of comparison of two frequency distributions sometimes desired. The question may be raised, for example, whether there is greater variation in the earnings of one group than another; whether carpenters' earnings vary more than those of common labor. The answer to this sort of question involves the spread of all the observations about the position of maximum frequency. Chart 23 furnishes an answer to this question for the distributions there shown — about as exact an answer as is possible with a graph; but the answer would not be so easy in the case of carpenters and common laborers if carpenters' earnings tended to concentrate around $60 per week and the other group around $20. In such a case the same spread of the two distributions on the graph (in numbers of dollars) might have a totally different significance in the two groups, the important thing being relative spread. That is, a $10 spread among common laborers' earnings might have the same significance as a $30 spread among carpenters' earnings. The case is illustrated by the data of table 13 and the graph of them on chart 25. Here the class of greatest frequency is, for carpenters, $50, and for common labor $20. To compare the relative spread of the two distributions on a chart, it is necessary to construct two horizontal scales, one for each distribution, and they must have a common origin at zero, the one case where a zero position on the horizontal scale of a fre-

TABLE 13.

WEEKLY EARNINGS CARPENTERS AND COMMON LABOR

WAGE CLASS	CARPENTERS		COMMON LABOR	
	Numbers	Percentage	Numbers	Percentage
15—	...	.....	65	14.13
20—	...	.....	143	31.09
25—	...	.....	118	25.65
30—	...	.....	87	18.91
35—	12	3.00	42	9.13
40—	47	11.75	5	1.09
45—	79	19.75	...	.....
50—	134	33.50	...	.....
55—	92	23.00	...	.....
60—	31	7.75	...	.....
65—	5	1.25	...	.....
Totals	400	100.00	460	100.00

quency graph is necessary. The scales are constructed so as to equate positions of equal significance in the two distributions. Generally the position of the two arithmetic means will be



equated, or of some other average value that is significant. In chart 25 the position, $25, on the one graph is equated with $50 on the other, which is approximately equivalent to equating the arithmetic means, since they are respectively $26.55 and $52. Though the table shows that the absolute spread of carpenters' earnings is greater than of common laborers' earnings, since they occupy seven classes compared to six; yet the comparison of relative spreads on the chart shows clearly a much greater relative variation in the earnings of common labor.

EXERCISES

I

Obtain from each member of the class an estimate of his or her weight and construct from these estimates a frequency distribution of weights with, say, a five-pound or a ten-pound class interval. Construct a histogram, a frequency polygon and a frequency curve to represent the data.

II

Data from the *United States Statistical Abstract*, 1928, page 328. *Index Numbers of Retail Prices of Foods in Principal Cities.*

Construct frequency distributions, with one- and two-unit intervals, of the index numbers for different cities and from these data construct histograms, frequency polygons, or frequency curves.

III

Data from *Monthly Bulletin* of United States Bureau of Labor Statistics.

The Bureau currently collects over five hundred separate quotations on wholesale prices of commodities and at frequent intervals publishes these prices and also price relatives in the *Monthly Bulletin.* Excellent use may be made of these lists of price relatives for making frequency distributions and their graphs and for thus comparing results for different periods of time or for different groups of commodities.

For example, compare the entire frequency distribution of price relatives 1926 and 1927; or compare farm products, 1927 with foods 1927, or with textiles or metals, etc. These comparisons will bring out

many of the characteristics of frequency distributions in economic data, such as skewness in the distributions, and the different shaped curves for different groups of price relatives — the lack of homogeneity prevailing between these groups.

IV

Comparisons of frequency distributions by graphs. Data from retail prices of foods (in II above), or from wholesale prices (in III above). Construct on one diagram the graphs of the two distributions to be compared. Try the histogram, polygon, and curve and select the one which seems to be most effective for the purpose.

V

Cumulative frequency distributions and graphs. Construct cumulative frequency distributions of the data of II and III and draw cumulative frequency curves to represent them. Comparisons of curves for food prices in two years, or for two commodity groups of wholesale prices can be made on the same chart.

(*c*) HISTORICAL GRAPHS

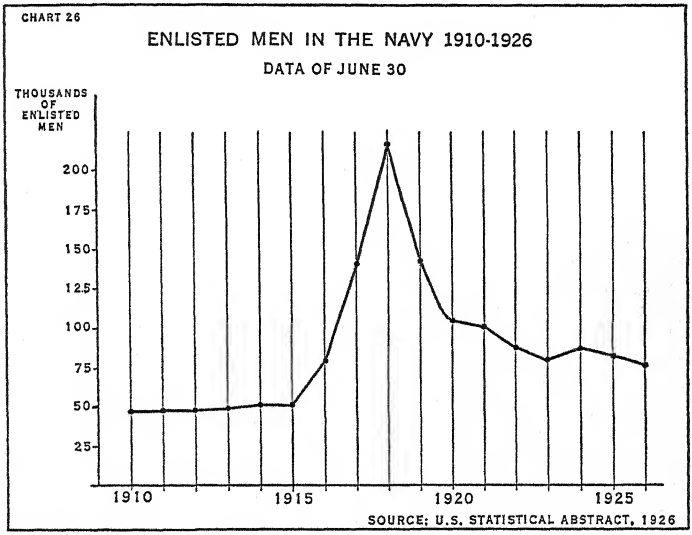
Basic graph for showing a statistical variate as a function of time. Historical series express the values of statistical variables for consecutive periods of time — hours, days, weeks, months, years, or any other subdivision. As previously observed, the fluctuations of an economic time series may for convenience be classified into four main types: trend, seasonal variations, cyclical variations, and irregular variations. The classification is based upon the idea that each type of fluctuation is the result of an assignable set of economic forces; and the methods of graphic statistics appropriate in any given case depend upon the type of fluctuation that is to be represented. The economic forces in question produce their effects through passing time, and it is but a short-cut, therefore, to say that a given variable is a function of time. The graphic method of representing the functional relationship of two variables in a plane is therefore the basic method of representation. It is illustrated for one of the simplest cases in chart 26, showing a

TABLE 14.

ENLISTED MEN IN THE NAVY
JUNE 30 OF EACH YEAR 1910–1926

YEAR	NUMBER ENLISTED	YEAR	NUMBER ENLISTED
1910	45,076	1919	145,018
1911	46,759	1920	107,360
1912	46,651	1921	103,571
1913	48,068	1922	88,580
1914	52,667	1923	82,355
1915	52,561	1924	87,327
1916	77,956	1925	84,289
1917	141,543	1926	82,161
1918	217,834		

Source: *U.S. Statistical Abstract*, 1926.



graph of the data in table 14. The time variable is placed upon the horizontal scale and the other variable upon the vertical.[1]

[1] Historical graphs are sometimes constructed with the time variable on the vertical scale but the practice is opposed by most students of graphic statistics and would better be avoided.

The vertical scale begins at zero and is set to bring the highest point of the graph near the top of the space to be devoted to the diagram. The abscissæ of the plotted points are marked by the years, that is, 1910, 1911,..... 1926. It should be recognized, however, that any year represents on the horizontal scale, not a point, but a distance along the scale equal to the distance between any two vertical lines as drawn since ordinates are shown on the diagram one year distance apart. The proper plot of a point for any year should be, for the abscissa, the mid-position of the distance represented by that year; the year designations on this graph, therefore, are placed at the mid-points of the year-intervals, and it should be recognized that the year 1910 extends on the horizontal scale a half-year distance on either side of the ordinate marked 1910; and similarly for the other years. The advantages of the designation as given on this graph lie in the greater ease with which the points may be plotted on the ordinates as drawn rather than at a position midway between two ordinates.

Of the three properties of this type of functional graph,[1] slope, continuity, and area, the first two are significant for the graph of enlisted men. The change in numbers of enlisted personnel from year to year is clearly indicated by the changing slope of the curve, especially the large increase from 1915 to 1918 and the fall from 1918 to 1920. The continuity of the curve likewise has real significance in this case; the data being for June 30 of each year, the line drawn between each two plotted points is therefore an approximation to the true, though unknown, numbers of enlisted men at other times of the year. No significance, however, attaches to the area under the curve, since the data do not represent a temporal distribution, and the numbers for the various years cannot therefore be added, or have no meaning when added.

There is one further property of this curve of which much

[1] See pages 99 to 101.

use is made in time graphs, namely, vertical distance as a measure of the magnitude of the dependent variable (time is considered the independent variable in an historical series). Frequently the main purpose of such a graph is to set forth the comparative values of a variable for the several time periods. Such questions as the following are suggested: How does the enlistment for 1918 compare with pre-war years? How do more recent post-war years compare with pre-war? These questions are answered by comparison of the heights of the curve at the periods in question. The ordinate of the graph at 1910 may be compared directly with that of 1925, or the series of ordinates 1910–1913 with those of 1922–1926. Not only is it possible to tell readily from the graph which of the two periods show the greater value of the dependent variable, but to a high degree of accuracy the ratio of the two can be estimated. This property of the curve, comparison of vertical distances, is analogous to the linear comparisons of magnitudes discussed on pages 62 to 67. In order to make this comparison on the time graph, it is most essential that the vertical scale begin at zero. If, in chart 26, for instance, the scale began with 25, the comparison of 1910 with 1926 would show for the later year a distance above the base line more than twice that for 1910, a result arising from cutting off 25 units from each ordinate; while the correct comparison of these two numbers by the full lengths of the ordinates from the zero position indicates that the 1926 figure is but slightly more than one and one half times that of 1910. When it is desired on a time graph to make these comparisons for different periods the zero position of the vertical scale must be shown.

Time series of aggregates — Bar graphs and curves. Chart 26 furnishes the basic method of representing a time series. Variations from it arise as a result of emphasizing different kinds of data or different types of fluctuations. The data of time series have been classified as (1) frequencies or aggregates,

(2) magnitudes or (3) derived data, such as averages, rates and ratios. When the data are of the first type, frequencies or aggregates, it is sometimes desired to indicate their character by the method of graphic representation. The case is best presented by contrast of two types of data: the case of enlisted men in the navy, shown in chart 26, furnishes one type, while yearly figures of pig-iron production illustrate the other. A continuous curve is the appropriate representation of the facts for enlisted men; for if full data were available — that is, a full record of the exact times at which enlistment changed either by discharge or by new enlistments — the continuous curve, fluctuating in accordance with these changing numbers, would correctly represent the facts. At any instant of time there is an enlistment total which is represented by an ordinate of the proper height for the abscissa of that time. But in the case of yearly figures of pig-iron production, each year begins with a production equal to zero and the figures continue to increase until the end of the year. A line connecting points representing these totals for the various years, each plotted at the midpoint of the year, therefore, does violence to the facts represented, for the continuous line indicates values of the dependent variable — that is, ordinates — at any subdivision of time, whereas the true representation of the facts requires introduction of the conception of *growing up* to a total in each subdivision of time. The distinction between these two conceptions is most important. It is the distinction between a thing that starts at zero in each time period, then grows up to a given total, and a thing which merely changes its level; it is the distinction between the growth of a tree and the changing level of the ocean. This idea of growing up is, of course, appropriate to most data of aggregates or frequencies, but not to all cases of them, as for example the population of the United States at different census periods. There is nothing in the nature of *magnitude* data, as the term has been

used here, to make inapplicable the conception of growing-up, but the fact remains that in economic data it does not ordinarily apply, for magnitude data are commonly of the fluctuating or level-changing type.

If the distinction between these two types is to be indicated in the graphic representation of a time series of aggregates, it is done by drawing an ordinate in each time interval of the diagram, but not connecting the tops of the ordinates, when the growing-up representation is to be made. While the ordinate may be merely a line, it is usually more satisfactory to erect bars of a width of half the time interval or more. The advantage of bars over lines lies in the greater prominence thus given to the graph of the data. Bars of a half-interval in width are usually to be recommended, for that gives ample space between them for differentiating one from another and emphasizes the growing-up conception. This type of graph is illustrated in chart 27, showing net imports of tea into the United States. This type of historical graph is to be used with data showing distribution of frequencies or aggregates in time when the peculiar character of the data is being emphasized. In many in-

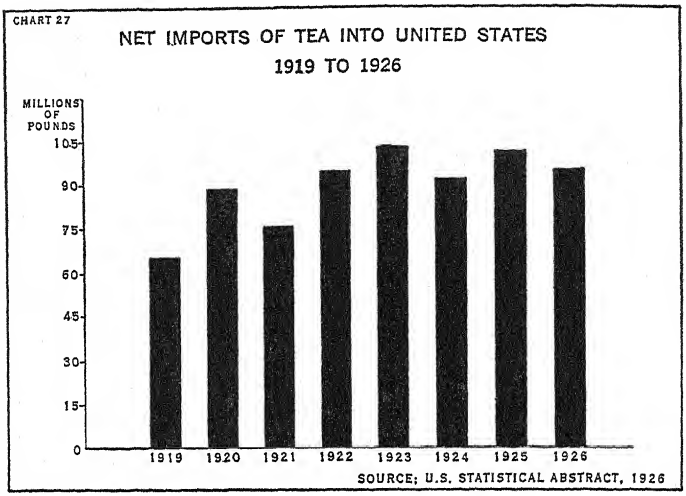
TABLE 15.

NET IMPORTS OF TEA INTO UNITED STATES 1919 TO 1926

UNIT: 1000 lbs.

YEAR	NET IMPORTS
1919	65,074
1920	87,801
1921	75,002
1922	93,888
1923	102,157
1924	90,496
1925	99,467
1926	94,512

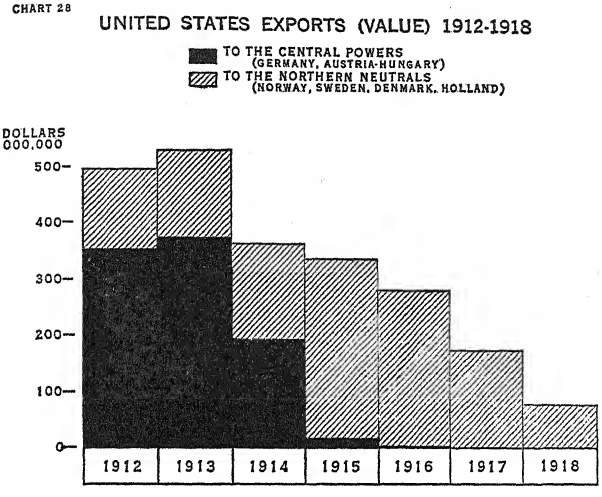
Source: *U.S. Statistical Abstract*, 1926



stances where a temporal distribution is involved, the interest in graphic representation lies in another aspect of the data, namely, in comparison of the changing totals of the successive time periods, and when this is the case, the continuous curve connecting the tops of ordinates is the proper graph to use. Chart 28, reproduced from the *Report of the War Trade Board*, 1920, a unit in the war-time organization at Washington, furnishes an excellent illustration of the historical bar chart and of the success that sometimes follows consideration of several methods of representing the data. The purpose of this chart was to present the data of United States exports to the Scandinavian countries, which remained neutral during the World War, and of United States exports to the Central Powers, in order to discover whether our exports to the former were sifting through their borders to Germany and Austria-Hungary while we were neutral and whether this sifting process ceased after we became a belligerent. The data for this chart were obtained by months from 1912 to 1918 inclusive and were plotted

CHART 28

UNITED STATES EXPORTS (VALUE) 1912-1918



SOURCE: WAR TRADE BOARD REPORT, WASHINGTON, D.C., 1926, P. 31

as an historical bar chart by months. But the monthly fluctuations were so violent as to make it quite impossible to trace the possible movement of goods from the neutrals to the Central Powers; next the data were combined into quarterly periods and plotted, but the fluctuations from period to period were still so great as to make difficult any decision on the question of reëxports to the Central Powers. Finally yearly data were plotted, as given in the reproduction, the bars for each year representing total United States exports both to the Central Powers and to the neutrals, the former in solid black, the latter shaded with diagonal lines. The conclusion now stood out clearly on the chart. In the pre-war years 1912 and 1913 exports to Germany and Austria-Hungary were about twice the value of those to the Scandinavian countries. In 1914 the latter obtained about their usual amount from us, but, thanks to the British blockade, our exports to the Central Powers were

cut off almost entirely in the last half of the year. In 1915 we sent practically nothing directly to the Central European belligerents, but almost doubled our exports to the Scandinavian countries, and the result was much the same in 1916. The evidence, finally, that after we entered the war in 1917 these unusual exports to the northern countries ceased is unmistakable.

Graphs for showing fluctuations. When the purpose of a time graph is *comparison of the values of the variable* for different time periods, the basic graph is the proper one to use, whether the data be of frequency, of magnitude, or derived from one of these; or the basic graph as modified through the use of bars in place of a continuous line. Of the four types of fluctuations of a time series, the one of chief interest in the cases so far considered is trend. The comparison of the sizes of the variate at two time periods is essentially a trend comparison, although the term as usually used involves this sort of comparison through a sequence of periods. But for the other three types of fluctuations, namely, cyclical, seasonal, and irregular variations, the interest lies not so much in change of level in the variable as in the alternations around a given level, the ups and downs of the curve. The usual conception of the business cycle, as it affects various economic time series, is just this, that production is very active for a time, is above normal, and later slows down to the point where it is below normal; equally with price data, prices for a time reach a high point that they cannot maintain and they fall, but again they are likely to fall too far and will come back; and so with many other economic series that reflect the business cycle. A seasonal variation differs in no fundamental way from the type just described except for the difference in length of time involved and for the greater constancy of the period of the fluctuation. A seasonal movement is of course a movement that completes itself within a twelve-months period, and, except for slight variations that need not

be considered here, is an exact period of one year. Cyclical fluctuations are longer than a year and vary a great deal in length, sometimes two years, or three, or even four; and there may be cyclical fluctuations or "long waves" that require ten, fifteen, or twenty-five years, or even longer. But from the viewpoint of graphic representation all these fluctuating or cyclical movements of time series, and the so-called irregular movements as well, may be treated as alike. The purpose in graphic representation is to show one or both of two phases of the fluctuation, its amplitude or its period. The former has reference to the extent of the variation of the phenomenon from its normal and is indicated by the vertical distance through which the curve fluctuates. It is usually of less importance in economic time series than the period of the fluctuation, or the length of time required to complete an up-and-down movement. It is the portrayal of the latter that enables one to compare the fluctuating movements of one series with another and to establish relationships of precedence or of sequence between different economic phenomena — relationships that may exist within the various parts of a business or an industry or that may be discovered between related industries. The discovery of these relationships is an important task of the business man and the economist and it is this fact that gives to the graphic representation of the fluctuations of economic time series its great importance.

The graphic representation of the fluctuations of a time series is accomplished by a slight modification of the basic time graph, through adjustment of the vertical scale. The interest in the alternating movements of the series requires that these movements be made prominent, whereas when the trend or level of the series is to be emphasized, they may be so small as to appear insignificant in comparison with the trend. This result is accomplished by showing, not the entire vertical scale from zero to the highest value involved, but just enough of the scale to

bring into prominence the fluctuations. Charts 29 and 30 illustrate the two cases with the same series, in the first of which, chart 29, the interest lies in the total magnitudes involved — that is, the height of the curve above the zero position; while in chart 30 the height of the curve is of no interest, but its fluctuations are emphasized and the vertical scale is shown only between the limits 120 and 180. The fact that the vertical scale is not shown in its entirety should always be made clear on the graph, by the double saw-tooth line, as in chart 30, or by some equally appropriate device. Except for the difference in vertical scales, charts 29 and 30 are identical in construction and both the curves refer to the same data. The impressions gained from observing the two graphs, however, differ greatly. The first is of a slightly waving or fluctuating line that gradually increases its distance from the base; in the second the relationship to the base is deliberately obscured, but the fluctuating movement around the gradually increasing level of the series is made prominent.

TABLE 16.

AVERAGE DAILY FREIGHT CAR LOADINGS

ADJUSTED FOR SEASONAL VARIATION

Unit: 1000 tons

MONTH	1919	1920	1921	1922	1923	1924	1925	1926
Jan.	131.3	147.1	127.7	132.9	159.0	160.2	166.7	169.3
Feb.	127.7	144.0	126.8	142.2	157.0	167.3	167.3	170.0
Mch.	127.1	155.0	126.3	151.0	156.2	167.0	168.6	176.0
Apr.	129.4	135.2	128.0	132.6	170.2	160.1	169.7	174.9
May	133.2	145.2	129.0	134.5	169.3	156.8	170.5	175.1
June	134.1	153.8	130.0	143.2	168.1	149.5	163.9	177.0
July	145.3	153.7	128.7	141.4	169.2	152.0	167.2	178.8
Aug.	142.8	154.5	129.7	139.2	166.0	155.0	172.1	176.8
Sept.	148.8	149.7	130.3	144.8	161.2	161.1	167.0	177.3
Oct.	144.9	150.7	138.8	149.0	160.7	163.8	165.8	180.9
Nov.	143.0	148.9	127.4	157.0	163.3	164.9	172.4	181.0
Dec.	143.8	146.7	128.2	161.9	158.3	165.8	173.8	178.6

Source: Ellsworth: *The Annalist Index of Business Activity, Revised.* Reprinted from *The Annalist*, Jan. 28, 1927.

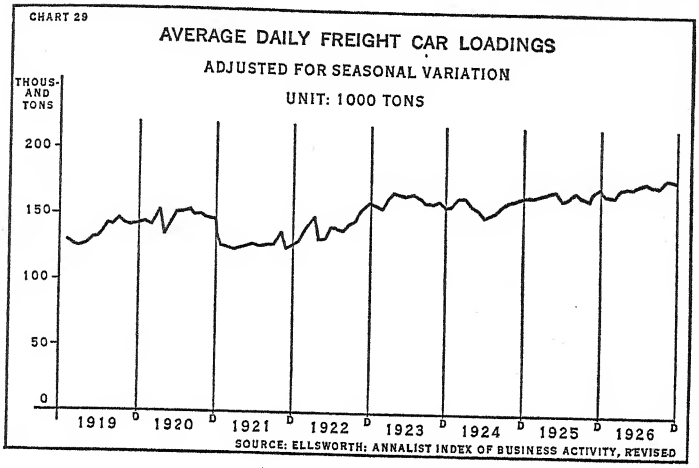


CHART 29
AVERAGE DAILY FREIGHT CAR LOADINGS
ADJUSTED FOR SEASONAL VARIATION
UNIT: 1000 TONS
THOUSAND TONS
200
150
100
50
0
1919 D 1920 D 1921 D 1922 D 1923 D 1924 D 1925 D 1926 D
SOURCE: ELLSWORTH: ANNALIST INDEX OF BUSINESS ACTIVITY, REVISED

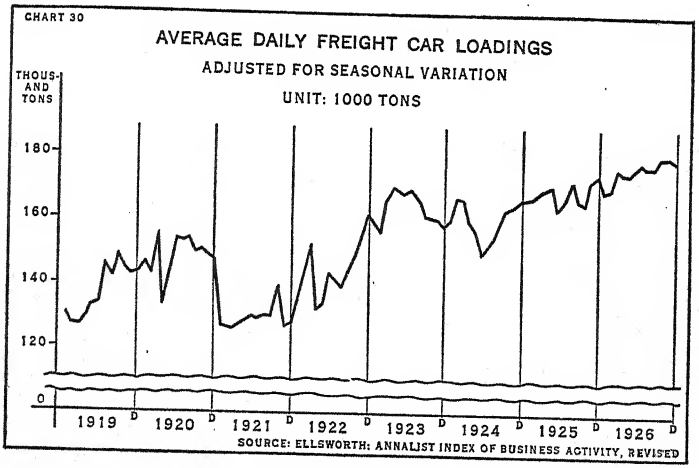


CHART 30
AVERAGE DAILY FREIGHT CAR LOADINGS
ADJUSTED FOR SEASONAL VARIATION
UNIT: 1000 TONS
THOUSAND TONS
180
160
140
120
0
1919 D 1920 D 1921 D 1922 D 1923 D 1924 D 1925 D 1926 D
SOURCE: ELLSWORTH: ANNALIST INDEX OF BUSINESS ACTIVITY, REVISED

The matter of deleting that portion of the vertical scale between zero and the smallest values of the variable, in order to accentuate the fluctuations on the graph, is especially important in those cases where the fluctuating movement is small in comparison with the size of the variable, where for instance fluctuations do not ordinarily occur of a value more than a few per cent above or below normal. If, however, fluctuations regularly occur of forty, fifty, or some such large percentage of normal values, they will be displayed adequately even though the vertical zero of the graph is shown. The special construction whereby the vertical zero is omitted is therefore necessary only when the fluctuations are small relatively to the size of the variable. A ten per cent fluctuation in some variables may have greater economic significance than a thirty per cent fluctuation in others, and it is this significance that is visualized by amplifying the vertical scale. To take but one illustration, the graphs of daily average pig-iron production or automobile production for monthly data show fluctuations very large in comparison with the trend of the data and these fluctuations are sufficiently emphasized when the vertical scale shows the full values from zero to the maximum, whereas this is not true of the figures of car loadings.

Cumulative time series and their graphs. Historical series of frequencies or aggregates when of the distributive type may be cumulated the same as the data of a frequency distribution of magnitudes and the method of showing such cumulative data graphically is similar to the case of the frequency distribution. Cumulation in time series, however, is usually restricted to relatively brief periods of time, not infrequently to a year. There are few occasions with economic data where a useful purpose is served by cumulation of data over periods of ten years or longer; on the other hand, the cumulation of data during a given fiscal or calendar year is an important device and is used effectively by many business firms. The circumstances

under which it is valuable are illustrated by a firm which sets a goal for the year's business, which is, say, equal to or a certain amount greater than the business of the previous year. This goal may be attained by equal accomplishments over each of the twelve months of the year, and such a record may be indicated, or such a standard set, by showing on a time graph for one year a diagonal line running from the lower right-hand corner, the vertical zero on January 1, to an ordinate on December 31, which represents the year's goal. If against this standard be plotted on the same chart the actual cumulative performance of the firm by days, weeks, or months as the case may require, the comparison furnishes a constant check-up of performance with schedule and a spur to lagging efforts when performance falls behind. Or, in place of the diagonal representing the standard, the graph may contain the full cumulative record of the previous year's business and the curve for the current year may then be drawn as the year progresses, giving at any date a direct comparison of the two years.

The only points that need to be noted in the construction of the cumulative graph are (1) that the vertical scale begins at zero and extends to a maximum which will bring the year's total to the top of the diagram, and (2) the curve begins at zero on January 1, the January total is plotted at the end of the January interval and not at the mid-point, and similarly for the cumulative totals of the other months. This agrees with the method of plotting a cumulative frequency distribution, where the time scale on the historical graph takes the place of the magnitude scale on the frequency graph. The procedure is illustrated in chart 31, a redrawing of a graph in the *Report of the War Trade Board.*[1] The data are given in table 17. The purpose of this graph was to show the effects of the import restriction policy of the War Trade Board, established for the purpose in part of obtaining needed imports and at the same

[1] Government Printing Office, Washington, D.C., 1920.

time conserving shipping space. The policy was established of requiring imports to enter in their most concentrated form. Quebracho wood has about five times the bulk of the extract, hence restriction was imposed upon the import of this important tanning material in its bulky form, but imports of the extract were freely allowed. The restriction was placed in May, 1918. The 1917 cumulative curves for both wood and extract show the situation that existed just prior to the imposition of the restriction policy and the effect of this policy is shown by comparison of the 1918 curves with those of 1917. This comparison shows that the imports of the wood were fairly comparable for the first four months of the two years; but whereas the usual thing as indicated in 1917 was for most of this material to arrive in the more bulky form, in 1918 as a result of the import restriction the quantity of extract imported greatly in-

TABLE 17.

UNITED STATES IMPORTS OF QUEBRACHO WOOD AND QUEBRACHO EXTRACT

CUMULATIVE TOTALS BY MONTHS 1917 AND 1918

In Pounds

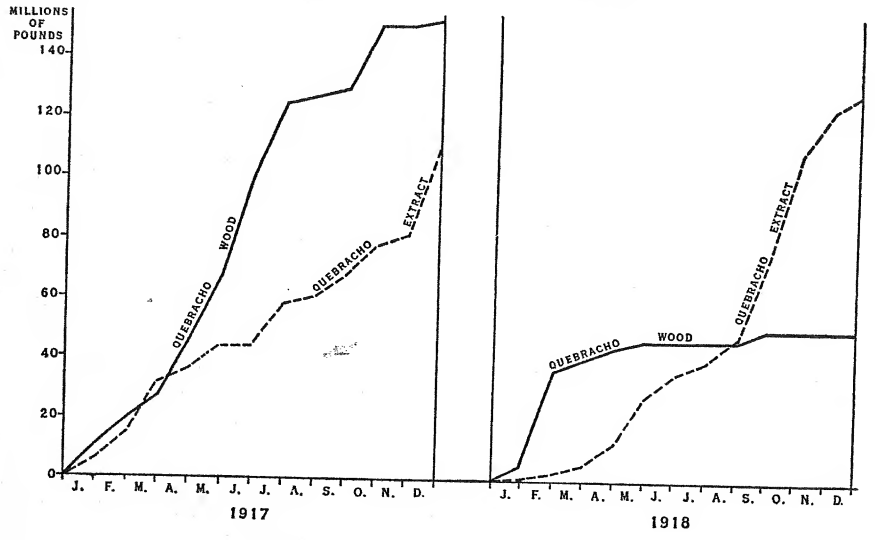
MONTH	TOTAL IMPORTS TO END OF MONTH			
	1917		1918	
	Quebracho Wood	Quebracho Extract	Quebracho Wood	Quebracho Extract
Jan.	5,992,000	5,775,892	4,872,000	420,785
Feb.	17,951,360	13,451,344	37,018,240	2,700,970
Mch.	25,553,920	31,637,995	41,027,840	5,027,270
Apr.	43,859,200	36,181,715	45,111,360	12,370,016
May	65,340,800	44,266,990	47,705,280	28,049,476
June	99,565,760	44,266,990	47,705,280	36,797,195
July	124,113,920	58,696,728	47,705,280	40,095,261
Aug.	126,965,440	60,477,095	47,705,280	48,070,898
Sept.	128,936,640	68,219,429	51,076,480	76,160,280
Oct.	150,230,080	78,358,387	51,076,480	109,196,189
Nov.	150,230,080	80,844,679	51,076,480	124,534,659
Dec.	153,646,080	108,993,077	51,076,480	131,109,739

Source: Report of War Trade Board, Washington, D.C., 1920.

CHART 31

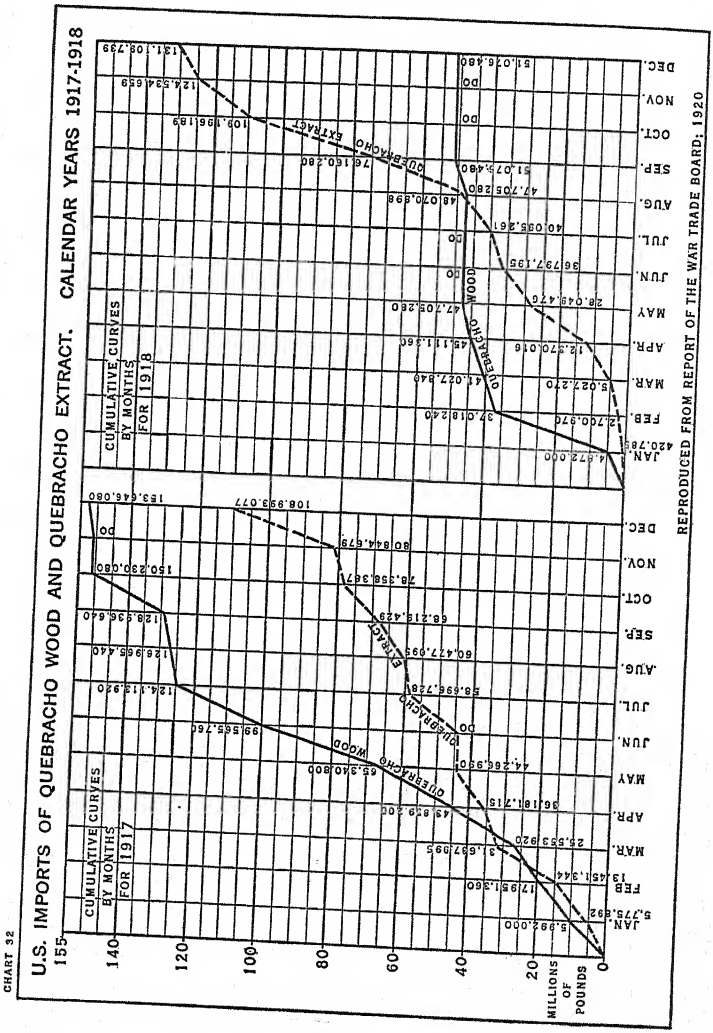
UNITED STATES IMPORTS OF QUEBRACHO WOOD AND QUEBRACHO EXTRACT

CUMULATIVE CURVES FOR 1917 AND 1918 BY MONTHS



SOURCE: REPORT OF WAR TRADE BOARD, WASHINGTON. D.C. 1920. P. 128

CHART 32


U.S. IMPORTS OF QUEBRACHO WOOD AND QUEBRACHO EXTRACT. CALENDAR YEARS 1917-1918
CUMULATIVE CURVES BY MONTHS FOR 1917
CUMULATIVE CURVES BY MONTHS FOR 1918
QUEBRACHO WOOD
QUEBRACHO EXTRACT
MILLIONS OF POUNDS
JAN.
FEB.
MAR.
APR.
MAY
JUN.
JUL.
AUG.
SEP.
OCT.
NOV.
DEC.
REPRODUCED FROM REPORT OF THE WAR TRADE BOARD, 1920

creased while the wood imports practically ceased after May. The chart bears testimony to the full effectiveness of the import restriction policy.

The original graph, of which chart 31 is a redrawing, is reproduced in chart 32 for the purpose of calling attention to a few details in which the original erred. The specification "calendar years" in the title is unnecessary, for the time scale of the graphs leaves no doubt about the question. The designation of the vertical scale is preferably placed above the scale figures rather than at the bottom. The month designations on the horizontal scale are incorrectly placed at the end of the month interval rather than at the mid-position. It is possible that the graph would be more distinct if there were fewer cross-section lines on the chart. The most serious defect of the chart by far lies in including alongside the curve the actual figures for each plotted point. A comparison of the redrawn chart with the original makes evident the much greater prominence of the curves on the former and the more successful result therefore in the presentation of the essential facts. The figures in the original graph should have been given in a table separate from the chart.

Graphs for comparisons of historical series. One of the valuable uses of historical graphs is in making comparisons of related series. The series compared may be of the simple or cumulative form, they may involve comparison of a whole and its parts or they may be entirely different though related series, and may be expressed in the same or in different units. The problem of comparison involves nothing new so far as concerns the construction of the graph for each individual series, but arises through the necessity of showing the two or more series on the same chart in such a way as to make important relationships between them stand forth. The question how many curves may be shown on a single chart may be treated first. The number should not be so great as to create confusion

through the multiplicity of lines or through frequent crossing of different curves. While no fixed rule is possible, it is certain that two or three curves may always be put on the same diagram, but when the number extends beyond four or five, better results will be obtained generally by using a second chart for some of the curves.

In general, the considerations governing the distinction between continuous and discontinuous data, between growing-up and changing levels, or between fluctuations and size, hold for comparisons of series as well as where a single series only is being shown. The problem of comparison, therefore, is the problem solely of bringing two curves into proper relationship on the same chart.

When the comparison is of a whole and its parts, the vertical scale of the diagram is determined by the requirements of the curve of totals. It will ordinarily be plotted the same as though it were the only curve on the chart, or where possible it may extend somewhat nearer the top of the diagram than when it is the sole curve. The parts, if there are not more than, say, two or three, may be plotted to the same vertical scale and with the same base line, as is done in chart 33. This permits comparison of sizes of corresponding ordinates of the several curves and of times and directions of change in the several curves. The same diagram may be used to illustrate another way of charting the whole and its parts in a time series. If, for instance, the curve for closed cars had been omitted and the area under the *open cars* curve shaded, then the area between this curve and that for total cars shaded in a different manner, the result would be a whole-and-its-parts comparison analogous to the subdivided bar.[1] This method shows how the total for each time subdivision is broken up into its parts, and the changing widths of the various *zones* indicate the changing sizes of the several parts.

[1] See page 79 ff.

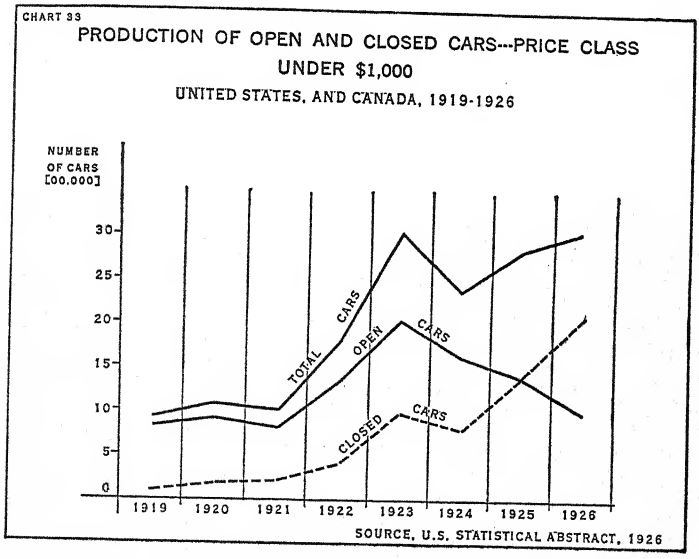
TABLE 18.

PRODUCTION OF OPEN AND CLOSED CARS — PRICE CLASS UNDER $1000
UNITED STATES AND CANADA, 1919–1926.

YEAR	TOTAL NUMBER	NUMBER OPEN CARS	NUMBER CLOSED CARS
1919	976,400	888,500	87,900
1920	1,118,600	904,900	213,700
1921	1,044,700	820,100	224,600
1922	1,774,400	1,343,000	431,400
1923	3,034,900	2,052,800	982,100
1924	2,391,900	1,622,800	769,100
1925	2,808,500	1,409,100	1,399,400
1926	3,025,700	967,500	2,058,200

Source: *U.S. Statistical Abstract*, 1926.

CHART 33



The data of this diagram could equally well have been represented by vertical bars erected in each time subdivision and then each bar could have been subdivided into appropriate parts representing open and closed cars. The effect of each method, bars and curves, is to show zones one above the other, each representing a given subdivision of the variable. Chart 28 offers another illustration of a time graph showing the whole and its parts by different shadings of the subdivisions of vertical bars.

In charts 28 and 33 the data compared are expressed in the same units — dollars of exports in one case and numbers of cars in the other. Frequently the comparison is of series expressed in different units such as production and price, or bank reserves and interest rates. In such cases there is no possibility of comparing actual sizes of the two variables; but if at some time a normal relationship may be established between the sizes of two such series, it is then possible to show how over time the series vary from this normal relationship. One method is to calculate from each series a series of relative figures, representing the base of the comparison in each case as 100 and expressing all other items in each series as figures relative to this base. The two curves may then be plotted to a single vertical scale of relatives. The other method attains the same result by constructing separate vertical scales for the two series in such a way that the *normal* values of the two are represented by the same position on the vertical scales. The latter procedure is illustrated by the data of table 19 and chart 34. The two vertical scales of the chart were set so that the two curves began at about the same vertical height, the position 100 on the income scale being equated to the position 75 on the debits scale. This has brought the two curves into juxtaposition in such a way that they can be compared easily as to dates and direction of change. It has not seemed desirable to equate them exactly at any time within the period shown, for

TABLE 19.

DEBITS TO INDIVIDUAL ACCOUNT, MINNEAPOLIS FEDERAL RESERVE DISTRICT AND GROSS CASH INCOME OF MINNESOTA FARMERS 1919–1926

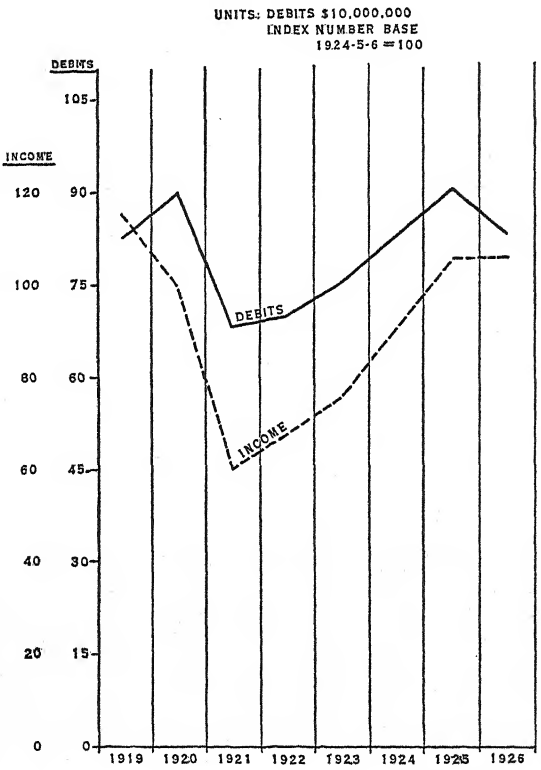
Index of Incomes: 1924–5–6 Base = 100
Debits in Millions of Dollars

YEAR	DEBITS TO INDIVIDUAL ACCOUNT	GROSS CASH INCOME OF FARMERS
1919	8,240	115.7
1920	8,902	100.3
1921	6,788	60.6
1922	6,971	67.8
1923	7,495	75.8
1924	8,240	90.4
1925	9,039	104.5
1926	8,301	105.1

Sources:
Income from Black and Kittredge, in *Jour. Farm Econ.*, July, 1928.
Debits from *U.S. Statistical Abstract*, 1926.

CHART 34

DEBITS TO INDIVIDUAL ACCOUNT,
MINNEAPOLIS FEDERAL RESERVE DISTRICT.
AND GROSS CASH INCOME OF MINNESOTA FARMERS
1919 - 1926



SOURCES; INCOME FROM BLACK AND KITTREDGE IN JOUR. FARM ECON. JULY 1928
DEBITS FROM U.S. STATISTICAL ABSTRACT, 1926

there is no certainty that any one year shows a more *normal* relationship between them than the others. In such cases it is generally sufficient to set the vertical scales so that the two curves lie fairly close together throughout their entire course. This is sometimes done by equating approximately the average values of the two series.

If it be desired merely to compare the fluctuations of two series to determine how nearly their maxima and minima coincide, and if there is no intention of comparing the relative amplitude of these fluctuations, then the vertical scales may be so set that the curves are brought closely together and so that the fluctuations are large enough to be prominent on the diagram; and it will not be necessary in this case that the two scales have the same zero position. If, however, it be desired to compare relative amplitudes of fluctuations of the two curves as well as dates of maxima and minima, then the curves must lie fairly closely together and their scales *must* have a common zero position.

The phenomenon of relative change. Adjustment of the vertical scales of the two curves on the same chart for the purposes indicated in the last section brings into prominence a feature of time series that is of great importance in economic data, the matter of relative or percentage change. When two curves are drawn on the same diagram, equating the vertical scales of the two from zero to the average or from zero to the magnitudes at a selected *normal* period, the purpose is to compare their changes relatively to this normal or base period. The importance of *relative* comparisons in economic data extends, however, farther than the case of two series; it applies equally to the items of either series taken by itself. The great significance of relative change, or rate of change, in an economic series lies in the organic character of most economic data. The magnitude of the changes which take place is conditioned by the size of the variable which is changing. For instance, in 1810 there were

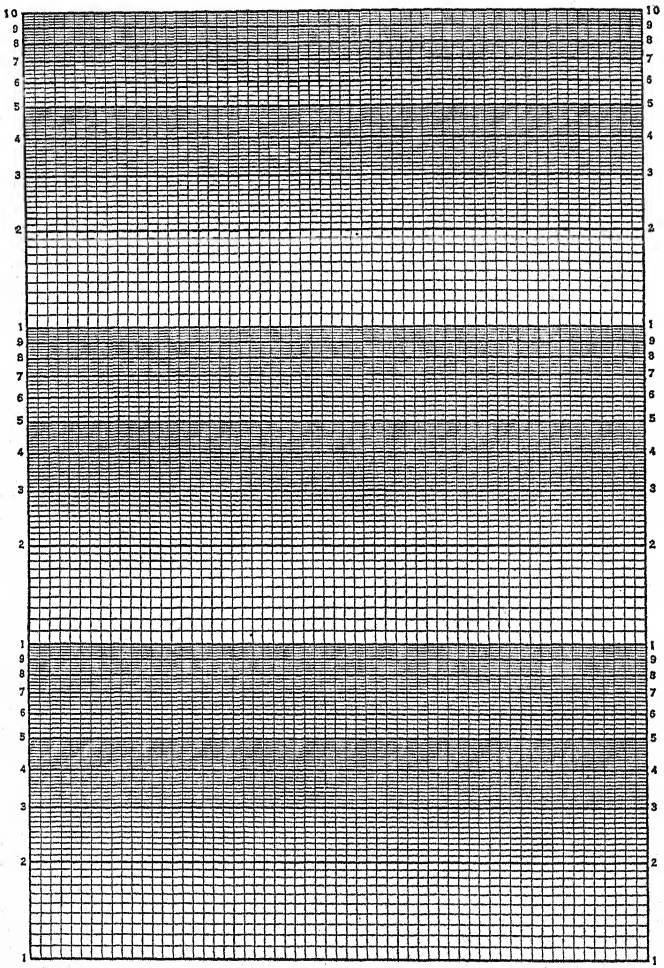
produced in the United States only 54,000 tons of pig iron; the figure had passed the 500,000 mark by 1850, while it has actually passed 40,000,000 tons since 1920. The difference between the 1920 production and that of 1921 was over 20,000,000 tons. This latter variation, of course, could not have occurred in the early history of the industry, and while it was due to the exceptional conditions of the post-war deflation, yet it is to be expected that the ordinary variations in the iron and steel industry, such as result from seasonal or cyclical swings, will be much greater nowadays than they were fifty or seventy-five years ago; that is, they are conditioned by the size of the industry at any given time. Fundamental to the changes that take place in many historical economic variables is the character of population growth. Economic variables grow or change in response to human demands for economic goods; and human population is an outstanding illustration of a phenomenon that grows in an organic way — its increment of growth over any period of time being conditioned by its size at the beginning of the period.

Semi-log, or ratio, charts. It is not the task of this book to delve more deeply into this peculiar character exhibited by much of the data of economics, but rather to take it as an accepted fact and to describe ways of showing relative change graphically. The device which serves this purpose is called the semi-logarithmic chart, a chart on which the time, or horizontal, scale is of the arithmetic type that has been used repeatedly in the previous pages, equal distances on this scale representing equal intervals of time; but the vertical scale is logarithmic. This means that equal vertical distances represent, not equal amounts or increments of change in the variable as has been the case heretofore, but equal rates of change. For illustration, on the usual graph drawn in rectangular coördinates, a vertical distance of one eighth inch may represent one hundred units, and this will be true whether the distance be selected at the top

or at the bottom of the diagram. But on the semi-log chart the same vertical distance represents, not a constant increment of change, but a constant rate of change. Thus one eighth inch may represent ten per cent change, so that from the point where the graph indicates a value of 100, this unit vertical distance will represent ten units of increment of the variable; but where the graph represents 10,000, the same vertical distance will represent 1000 units of the variable. This kind of chart takes its name from the fact that equal logarithmic distances — that is, equal distances on a logarithmic scale — represent equal rates of change in a variable. Such a situation can be shown on the ordinary type of graph if, instead of showing the actual values of the variable on the vertical scale, the logarithms of these values be shown, so that equal differences in the logarithms are represented by equal vertical distances. The conversion of the variable values into logarithms is, however, a routine task of some magnitude, and this work may be avoided by setting a logarithmic scale on the vertical of the diagram and plotting actual values of the dependent variable on this scale. Semi-log paper of this sort is now prepared and sold by a number of commercial firms.

An illustration of semi-log paper is given on page 148. This is known as a three-cycle chart. Scale figures are shown at the right and left of the chart running from 1 to 9 on each cycle. If the first cycle is taken to represent data figures varying from 1 to 10, the second cycle will then represent data from 10 to 100, and the third from 100 to 1000. The cycles are all of a constant width, indicating the fact that the logarithmic distance from 1 to 10 is the same as that from 10 to 100, or from 100 to 1000. If, again, the uppermost cycle on the chart represents variable values from 1 to 10, the middle cycle will represent values from one tenth to 1 and the lower cycle from one one-hundredth to one tenth. Thus it is seen that consecutive cycles differ by a factor of 10 and that there can be no zero posi-

SEMI-LOG PAPER - THREE CYCLES



tion on the vertical scale. The vertical scale can be used to plot values of any magnitude above zero, but is not appropriate for the representation of negative values.

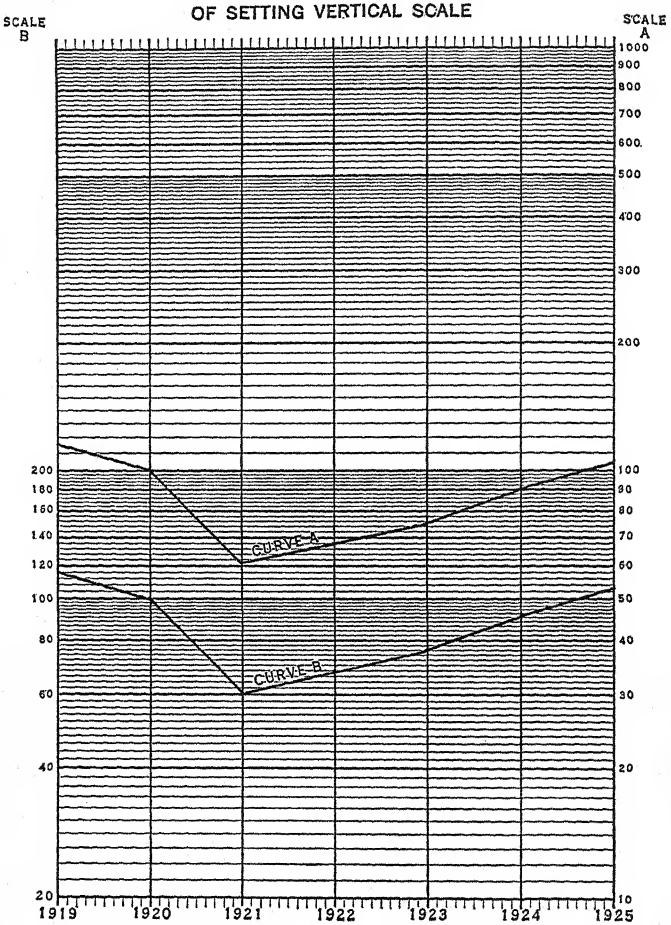
The plotting of a series on a semi-log chart differs in no way from the plot in rectangular coördinates, once the vertical scale figures have been set. The figures printed on the semi-log paper, as shown in the illustration, page 148, may be used and have been used for curve A of chart 35 by the addition of one cipher to each printed figure. The two curves of chart 35 are graphs of the indices of gross cash income, 1919 to 1925, in table 19. These index figures vary from 60.6 to 115.7, hence when plotted as in curve A will occupy space in both cycles of the chart. It is possible, however, to plot the same data on a single cycle chart, as is done with curve B, by setting a different scale on which the figures are each a constant multiple of the printed scale figures. The scale figures for curve B are each two times the size of the scale figures for curve A. The two curves lie at a constant vertical distance apart. One cycle is all that is needed to plot any series of figures of which the maximum is not more than ten times as great as the minimum; for a maximum greater than this but not more than one hundred times as great as the minimum, two cycles will be needed. Thus the total figures of table 18, varying from 9 to 30 (in hundreds of thousands) may be plotted on one cycle, the data of table 16 the same; but automobile registration figures, being 8000 in 1900 and 22,001,393 [1] in 1926, will require five cycles, as follows

1,000–	10,000
10,000–	100,000
100,000–	1,000,000
1,000,000–	10,000,000
10,000,000–	100,000,000

[1] See *United States Statistical Abstract*, 1926.

CHART 35

ILLUSTRATING TWO METHODS OF SETTING VERTICAL SCALE



if the printed cycle figures are used, but four cycles if each of the above cycle figures are multiplied by any figure that will raise the first cycle minimum not above 8000 and the fourth cycle maximum figure at least above 22,001,393. The integral multipliers that will produce this result are 3, 4, 5, 6, 7, and 8. The above cycle limits, therefore, multiplied by any one of these six factors will give a vertical scale permitting the plot of automobile registrations, 1900–1926, in four cycles.

Interpretation of ratio curves. Before considering actual cases of economic data plotted on semi-log charts, it will be well to see how curves on such charts are to be interpreted. Table 20, herewith, gives five series of hypothetical data for consecutive intervals of time. In the first three series, A, B, and C, the rate of growth between consecutive periods is constant, while in series D and E the amount of growth per period is constant, and therefore the rate declines as the variable increases. A and B have the same rate of growth, each value of the variable being four times the preceding value, but

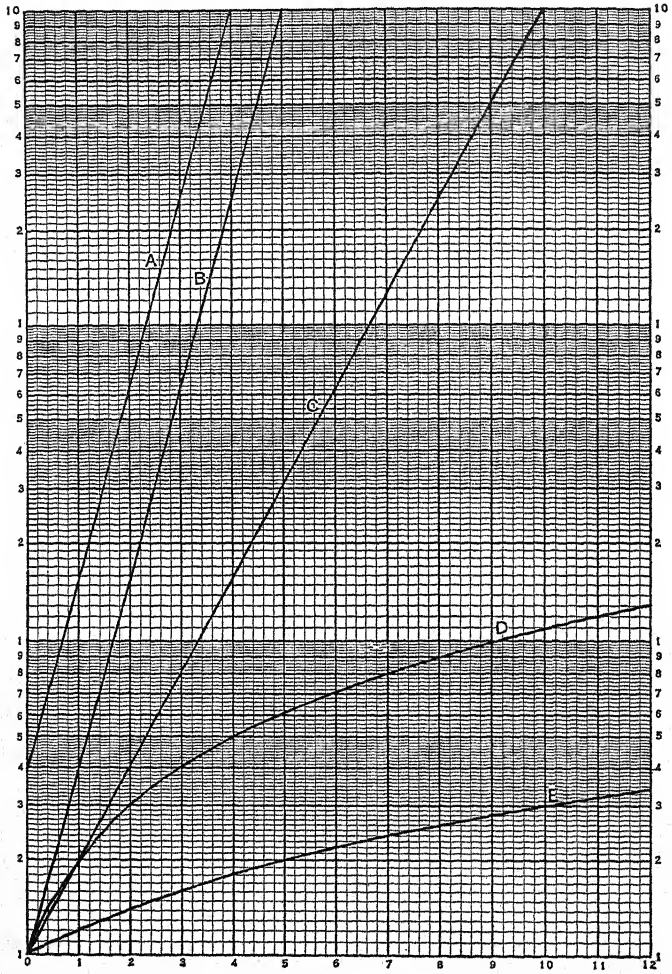
TABLE 20.

SERIES ILLUSTRATING CONSTANT AND DECREASING RATES OF GROWTH

TIME PERIOD	CONSTANT RATE SERIES			DECREASING RATE SERIES	
	A	B	C	D	E
0	4	1	1	1	1.0
1	16	4	2	2	1.2
2	64	16	4	3	1.4
3	256	64	8	4	1.6
4	1024	256	16	5	1.8
5	....	1024	32	6	2.0
6	....	....	64	7	2.2
7	....	....	128	8	2.4
8	....	....	256	9	2.6
9	....	....	512	10	2.8
10	....	....	1024	11	3.0
11	....	....	....	12	3.2
12	....	....	....	13	3.4

CHART 36

ILLUSTRATING CONSTANT AND DECREASING RATES OF GROWTH ON THE SEMI-LOG CHART

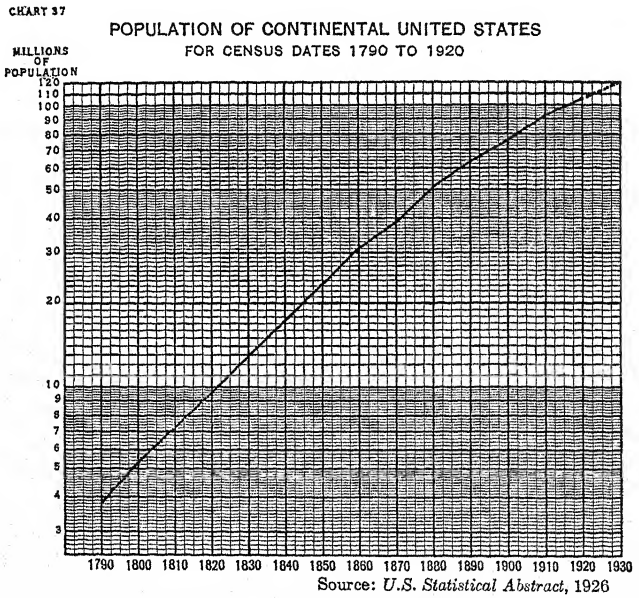


A starts at a higher value than B. The graphs of these two series on chart 36 show, first, that the two curves are straight lines, second, that they are parallel. The graph of C, which increases by a constant factor of two, is likewise a straight line, but is less steeply inclined than A and B. The graphs of D and E are curved lines with the concave side downward. These five curves illustrate several important characteristics of graphs on semi-log paper, which may be generalized as follows: (1) A constant rate of change is always indicated by a straight line on the diagram. (2) If the curve has a positive slope, the variable is increasing; if a negative slope, it is decreasing. (3) Differences in the slopes of the two curves or of two positions on the same curve indicate different rates of change and the steeper the slope the greater the rate of change. (4) Two curves with a constant vertical distance between them have the same rates of growth. The distance between them, if they are plotted to the same scale, indicates a difference in magnitude of the two variables at given times. This is true of course whether the graphs are straight or curved lines. (5) If a variable is increasing, but at a decreasing rate (curves D and E) the curve will be concave to the base or horizontal axis; conversely, if increasing at an increasing rate, it will be concave upwards. A decreasing variable with a declining rate of decrease will be concave upwards, while if it is declining at an increasing rate it will be concave downwards.

The ratio chart and economic trends. The semi-logarithmic, or ratio, chart is to be used when the purpose is to represent graphically relative change in a variable. Relative change is measured by the slope of the curve at any point. The points of more or less rapid change, either increase or decrease, are ascertained easily and with a high degree of accuracy with the eye since slight differences in the slope of a curve are easily detected. This statement is borne out by the graph of population in chart 37. The line is almost straight from 1790 to

1860, indicating an approximately constant rate of increase in the population of the United States between these two dates. The figures for percentage increase between census years are, for these dates, 35.1, 36.4, 33.5, 32.7, 35.9, and 35.6. The slight drop in the curve for 1870 is also clearly evident, a drop explained by the census authorities as due in part to incomplete enumeration in the Southern States. The tendency for the rate of growth to slow up since 1860 is also clearly apparent from the slight curvature of the line after that date.

One other valuable feature of the ratio chart is indicated by the population graph. Where the rate of change of a variable



Source: *U.S. Statistical Abstract*, 1926

is not constant, but changes very slowly in a given direction, it is frequently possible to extend the graph beyond the last date for which data are available and obtain thereby a fairly accu-

rate estimate of what the next figure will be. Such extrapolation can generally be performed more easily and more accurately on a semi-log graph than on an ordinary graph with arithmetic rulings, where the data are organic in character, for the rates of change of such a variable are likely to be steadier than the absolute amounts of change. The extrapolation of the population curve is indicated by the broken line from 1920 to 1930, showing an estimate of slightly over 120,000,000 by 1930.

The ratio chart and fluctuations in economic data. — Much emphasis has in recent years been given to the ratio chart for the purpose of plotting historical economic data, some of its exponents seeming to say that all economic time series should be plotted on ratio paper. This seems to put the case too strongly when it is recognized that sometimes comparison of absolute magnitudes is more to the point than comparison of rates of change. But it is true that there has been a greater tendency in recent years to emphasize the importance of relative change in economic data, particularly with reference to the alternating character of economic activity. If change is sufficiently gradual it may continue in a given direction for a long time — adjustments have time to be made; but if the pace of change is too great a reaction sets in. An important use of the ratio chart is thus to show these alternating movements of the business cycle. This is illustrated by charts 38 and 39, the former on ratio and the latter on arithmetic ruling. The data are for the monthly production of motor vehicles in the United States and Canada, 1922 to 1926 inclusive, as given in table 21. Consider first the upper curve in each chart, that for production of passenger cars. The question whether production is increasing so rapidly at any time as to bring a reaction can be considered in the light of what has happened in the past. Even over this short period of five years, annual production has increased so that a figure which represented an abnormal increase in activity in 1922 may not be abnormal for 1926. It is a question of rela-

TABLE 21.

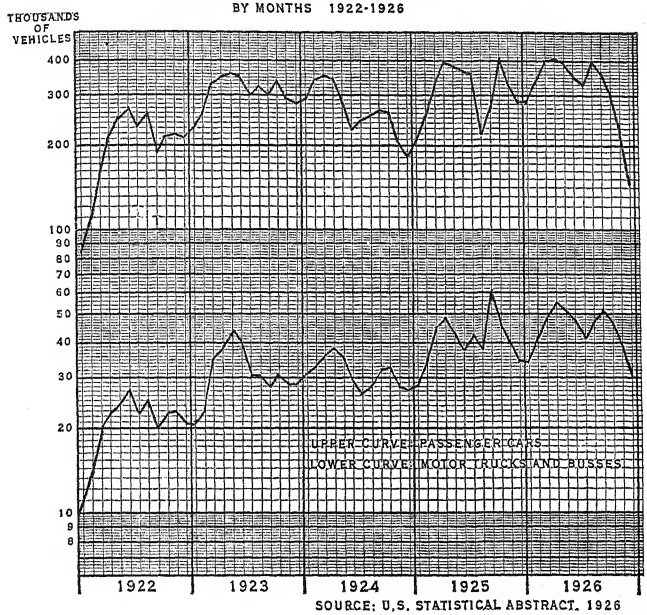
PRODUCTION OF MOTOR VEHICLES IN UNITED STATES AND CANADA

MONTH	NUMBER PASSENGER CARS					NUMBER MOTOR TRUCKS AND BUSSES				
	1922	1923	1924	1925	1926	1922	1923	1924	1925	1926
Total	2,397,827	3,719,164	3,262,764	3,835,801	3,929,546	251,434	378,288	378,107	498,470	535,197
Jan.	84,823	229,226	293,824	213,851	284,703	9,597	20,534	30,785	28,203	33,517
Feb.	111,843	260,881	343,460	253,955	334,524	13,455	23,143	32,974	34,482	41,784
Mch.	162,203	332,157	357,045	334,214	399,105	20,079	35,016	36,506	45,180	49,386
Apr.	208,543	349,474	346,405	393,262	401,836	22,613	38,640	38,037	47,984	54,135
May	244,634	360,743	286,324	384,548	394,569	24,293	44,125	35,408	43,719	51,568
June	263,127	346,059	225,079	366,510	358,388	27,030	40,639	29,135	38,151	47,265
July	230,554	305,795	244,544	360,124	329,959	22,636	30,139	26,448	41,870	41,873
Aug.	253,133	320,700	255,232	223,517	393,064	25,044	30,335	28,714	37,850	47,836
Sept.	191,156	304,087	263,528	274,227	363,547	20,258	28,100	32,015	60,482	51,257
Oct.	217,032	338,664	260,881	408,017	300,160	22,683	30,238	32,533	46,013	46,985
Nov.	218,200	289,553	204,343	337,435	226,278	22,813	28,639	27,956	40,048	39,430
Dec.	212,679	281,825	182,099	286,141	143,413	20,933	28,680	27,596	34,488	30,161

Source: *U.S. Statistical Abstract*, 1926.

CHART 38

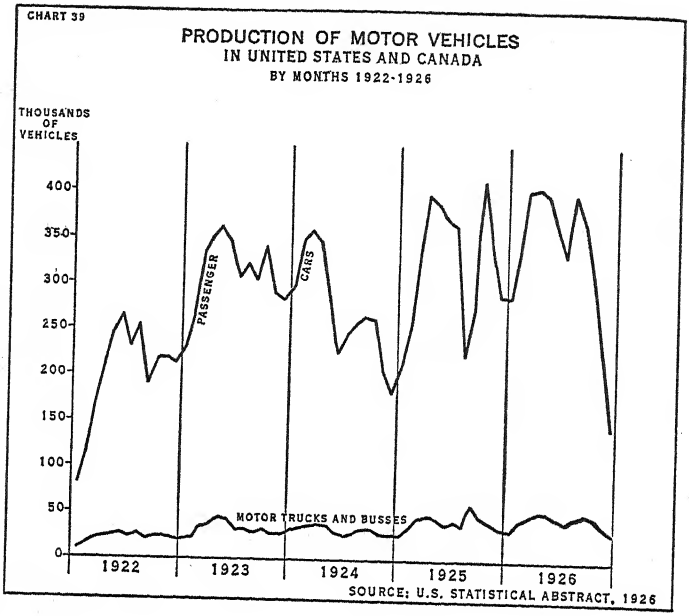
PRODUCTION OF MOTOR VEHICLES
IN UNITED STATES AND CANADA
BY MONTHS 1922-1926



SOURCE: U.S. STATISTICAL ABSTRACT. 1926

tive increases. From chart 39 relative increases for the first part of 1922 can be compared with those of the first part of 1925 only by reading approximate values from the chart and calculating percentages from these figures; on chart 38 the conclusion is reached immediately by comparison of the slopes of the two segments of the curve, in 1922 and in 1925. There is a close similarity in the rates of increase of these two segments, whereas in 1923 the rate of increase was slower. For this sort of comparison the ratio chart is to be preferred over the other.

One other important advantage of the ratio chart is revealed by the comparison of the two curves on the same chart. In each chart the two curves are plotted to a single scale. The



production figures for motor trucks and busses being so much smaller than those for passenger cars, the variations in the latter curve are not emphasized sufficiently in chart 39 to make any adequate comparison with the passenger car curve. Whether they fluctuate together or not is very difficult to say from this diagram. Of course it would have been possible to set a second scale on chart 39 so that the motor-truck curve would have been raised farther above the base line and have fluctuated within about the same amplitude as the other curve. But this necessity is avoided in chart 38 where both curves are drawn to a common scale, where the difference in their magnitudes results only in separating them on the chart and where the relative magnitudes of the two are directly comparable. The truck curve clearly shows a greater relative increase

(greater positive inclination) over the five-year periods than the other, while as to the shorter fluctuations of the two curves the amplitude is about the same. Direct comparisons of the slopes of the two curves is possible at any points for the determination of relative increases or decreases in the two series. In other cases, where the proper comparison of the two variables requires consideration of their actual sizes, the representation in rectangular coördinates is the correct one and cannot be replaced by the ratio chart.

EXERCISES

(*Note:* The *Survey of Current Business*, published by the United States Department of Commerce, is an excellent source for historical series of economic data. Some of the material found there is, of course, republished later in the *United States Statistical Abstract.*)

I

Data from *United States Statistical Abstract*, 1928, page 296, *Fire and Marine Insurance Business in the United States.*

Plot graphs representing for the years 1910–1926 (1) the number of stock and mutual companies; (2) their total assets; and (3) their total income. Are alternative methods available for drawing each of the three graphs?

II

Data from the *United States Statistical Abstract*, 1928, page 294, *Interest Rates.*

Construct a graph of the monthly fluctuations 1919–1925 in interest rates on six-months' loans, the purpose of the graph being to show whether there has been a cyclical fluctuation in interest rates during this period. Construct this graph first with a vertical scale beginning at zero, then by breaking the vertical scale and showing only sufficient scale to include the entire range of interest rates; and decide which method is the most effective presentation.

III

Comparison of fluctuations. Data of interest rates in II and of debits to individual account in *United States Statistical Abstract*, 1928, page 283.

Construct a graph to determine if there are cyclical fluctuations both in interest rates and in debits to individual account, and if the fluctuations of the two show any similarity. Both curves to be put on the same graph. Data monthly, 1920–1927.

IV

Comparison of historical series, simple and cumulative. Data from *United States Statistical Abstract*, 1928, page 446, *Merchandise Exports and Imports.*

(1) Draw two curves on the same chart, one for 1926 and one for 1927, to compare monthly values of exports (or imports). Are there alternative methods and if so is there a basis for choice between them?

(2) Draw two curves on one chart, one for 1926 and one for 1927, to compare cumulative monthly values of exports (or imports).

(Much use is made of this sort of comparison of one year with another, for both monthly and weekly data and on both a cumulative and a non-cumulative basis.)

V

Comparison of arithmetic and ratio charts. Data from *United States Statistical Abstract*, 1928, page 449.

Plot data of yearly United States exports of merchandise 1916 to 1927 on both arithmetic and ratio charts. Which is the most effective presentation of the figures? Why?

VI

Comparison of curves on ratio charts. Data from *United States Statistical Abstract*, 1928, page 476.

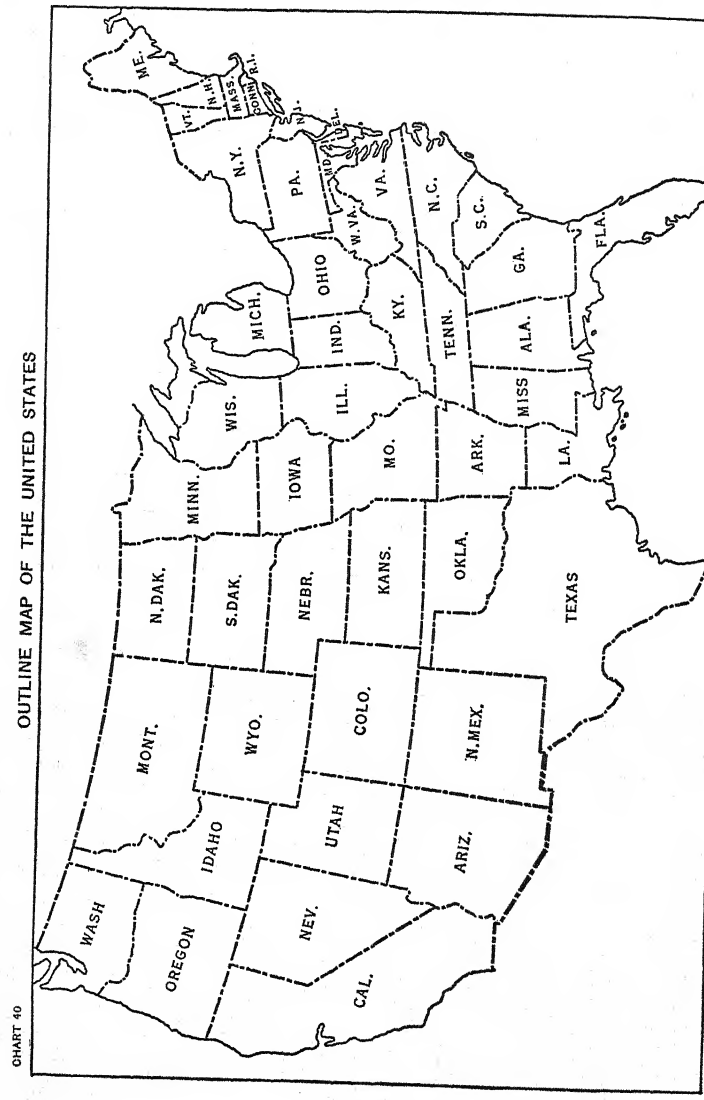
Plot two curves on a ratio chart to compare the trend and fluctuations in value imports of sugar and of burlap into the United States, 1916 to 1927.

IV

GRAPHIC REPRESENTATION OF GEOGRAPHIC DATA
STATISTICAL MAPS

GEOGRAPHIC classifications are peculiarly in need of a method of graphic representation that will throw into relief their spatial aspect since there is no way of presenting geographic data in tabular form that gives a satisfactory conception of the spatial distribution of the data. Geographic data can be shown, of course, on a bar chart or a circle and sector diagram, but this result focuses attention on the comparison of magnitudes one with the other or with a total and conceals wholly the aspect of distribution or variation of the data in space. But this very element of space comparison becomes in many instances the thing of greatest significance. Statistical maps are designed with the purpose of giving prominence to the spatial characteristic of the data.

The data of statistical maps. A geographic classification gives statistical data for subdivisions of a given area; it may be States of the United States, counties in a State, wards in a city; the subdivisions may be, and often are, non-political — sanitary districts in a city, sales districts for a business firm, or telephone exchange districts. The beginning of the statistical map is an outline map of the total area with boundary lines showing the area subdivisions. Ordinarily the outline map should contain very little other than these boundary lines, for the inclusion of rivers, cities, streets, or other elements not required in the presentation of the necessary statistical facts tends to divert the attention from the statistical facts themselves. Chart 40 shows a good outline map, the United States divided into States. The problem of graphic representation of spatial data involves the selection of appropriate graphic


OUTLINE MAP OF THE UNITED STATES
CHART 40
ME.
VT.
N.H.
MASS.
CONN.
R.I.
N.J.
DEL.
MD.
N.Y.
PA.
W. VA.
VA.
N.C.
S.C.
GA.
FLA.
OHIO
KY.
TENN.
ALA.
MISS.
LA.
MICH.
IND.
ILL.
WIS.
IOWA
MO.
ARK.
TEXAS
MINN.
N.DAK.
S.DAK.
NEBR.
KANS.
OKLA.
MONT.
WYO.
COLO.
N.MEX.
IDAHO
UTAH
ARIZ.
WASH
OREGON
NEV.
CAL.

methods of indicating statistical facts within each of these subdivisions of the map. These facts may be data of frequency or of magnitude.[1] Concretely, wheat may be the subject of study and, of the total United States product, given quantities are grown in each State. In technical terms, these are data of frequency — the frequency with which bushels, or other units, of the total crop are produced in each State of the total area. The series, or classification, is called a spatial distribution.[2] Again with wheat still the subject of study, the data for each State may refer to relative frequency; relative not to the total production of wheat; that is, not a component part relationship as that term has been used in the previous pages, but relative to some other phenomenon with which wheat production is necessarily associated — wheat production per acre or per capita population. These rates, ratios or averages, calculated for spatial categories such as States of the United States are obtained by relating one set of frequencies to another, as illustrated above — bushels of wheat to acres of land or to numbers of population. They are termed data of relative frequency or rates and ratios, and in their graphic representation are treated the same as magnitude data. The latter, degrees of a measurable quality or characteristic, require for their evaluation a scale of reference such as inches or pounds or degrees. Thus, heights of persons, lengths of leaves, prices of wheat, degrees of hardness of wheat, are magnitudes in the sense in which the term is used here. On the other hand, population per square mile, bushels of wheat per acre, and per cent of farms improved are ratios, but they are by nature analogous to magni-

[1] See page 70, where the distinction was first drawn between a frequency or aggregate and a measurable magnitude. The word *magnitude* is required to play a double rôle, for in the discussion of statistical maps it is used to refer specifically to a measurable characteristic, whereas in the discussion of "magnitude comparisons," it was used more broadly to cover not only the case of measurable characteristics but frequencies or aggregates as well.

[2] See Day: *Statistical Analysis*, chapter 7.

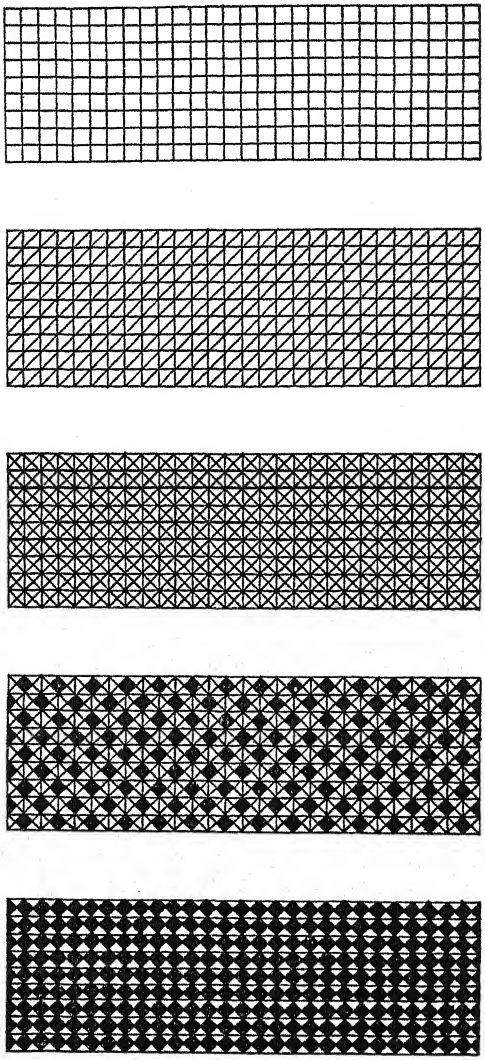
tudes and in their graphic representation are treated like magnitudes.

Methods of representing the data graphically. Graphic methods then are needed which will serve as adequate representations of these several types of data in the spatial subdivisions of an outline map. The geometric measurements heretofore used for showing magnitudes, namely lines and angles, are not suitable for this purpose, for to use them would require, for a United States map for instance, a display of forty-eight separate bars or angles in forty-eight locations, one in each State; and the resulting figure would give a confused and ineffective representation of the facts, little better than inserting the actual figures in each State. Circles or squares or solid figures, either spherical or cubic in shape, could likewise be drawn in each State to represent the magnitude or frequency involved, but these representations again are properly excluded because of the difficulty of estimating comparative magnitudes by means of them; and for a more pertinent reason, that, in this case of geographic representation, they do not give an impression that is appropriate to the data involved. For data of spatial distributions, or frequencies within geographic areas, the essential need is a clear and concise picture of the spread of these frequencies over space, a sense of relative densities; to use again a wheat illustration, it is as though a farmer sowed one bushel of wheat over one acre and two bushels over another. If now the two acres were represented by two spaces of similar size on an outline map and the appropriate number of grains of wheat shown on each space, there being twice as many on one space as the other, the representation would give to the observer an impression of greater frequency, or greater density, in the one case than in the other. And this picture would create a clearer idea of the facts than that obtained by placing in one area some such figure as 5,284,693 (grains or bushels or other units of wheat) and in the other area the figure 10,569,386 or twice the former number.

The illustration furnishes a clue to the possibility of representing such frequencies graphically in a spatial distribution. Each frequency, or each unit, may be represented by a dot or point; or if the number of frequencies is too great for this, the number of dots or points can be made proportional to the number of frequencies. With 500,000,000 bushels of grain to be represented, it is obviously out of the question to plot 500,000,000 dots, one to a bushel; but a dot may represent 10,000 or 100,000 bushels. On the choice of the number of frequencies to be represented by each dot two alternatives are presented, based upon different purposes in the graphic display. The unit may be made so small that there will be a very large number of dots — enough so that the densest areas of the map will show a practically continuous succession of dots, appearing as almost solid black in the finished map; or the unit may be made so large that there will be a relatively small number of dots, so small that they could be counted without great difficulty. These alternative methods may be called the *large-dot* map and the *point-dot* map.

The display of relative frequencies, or rates and ratios, on an outline map cannot be accomplished properly by the methods used for the spatial distribution of frequencies above, for the impression of density which is characteristic of the latter is not a true conception for the former. Relative frequencies are to be classed rather as magnitudes, and it is a magnitude conception rather than a density conception that is to be conveyed. But it has already been found that the usual graphic methods of showing magnitudes (the several sorts of geometric measurements) are not appropriate to the statistical map. The conclusion is unavoidable that no satisfactory method is available for showing comparisons of the exact values of rate or ratio magnitudes in the several spatial subdivisions of the map. But the matter does not end here, for it is possible to group the various rates or ratios into a small number of classes and to

CHART 41. DENSITY DISTINCTIONS FOR CROSS-HATCHING



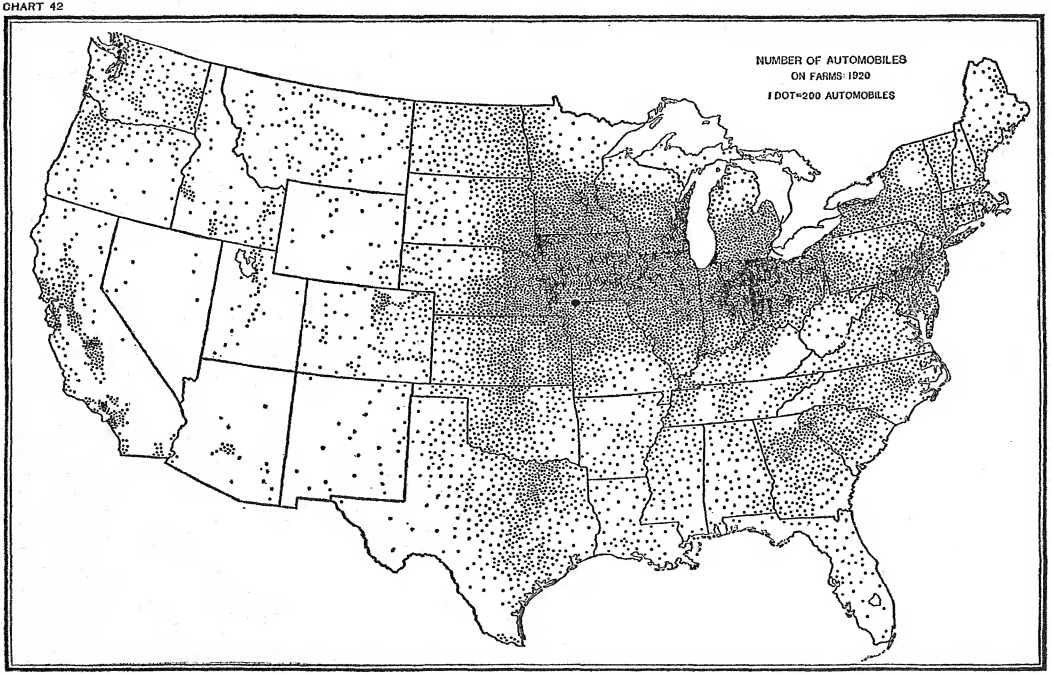
indicate the magnitude group or class into which each spatial subdivision falls by appropriate methods of cross-hatching. A system of cross-hatching can be constructed that will indicate clearly six, eight, ten, or even a dozen different degrees of magnitude and some authorities propose as many as twenty-five.[1] Given then a series of cross-hatchings, of which the lightest represents the smallest magnitude class and the heaviest the largest magnitude class and with proper gradations between these extremes, the several classes of magnitude of the rates or ratios may be indicated on the map by the appropriate cross-hatching on each subdivision and the result, if properly executed, will give a concise picture over the entire map of the relative size classes into which its different parts fall.

When it comes to the third type of statistical data shown for space categories, such as price, amount of rainfall, days of sunshine, and the like, the problem of graphic representation of these facts cannot be solved satisfactorily unless the magnitudes may be grouped in classes as was done with the ratio data above; in many instances this grouping is not satisfactory and if so the attempt at graphic formulation should be abandoned. If the grouping into classes can be resorted to, the method of cross-hatching may be used, the same as with classes of relative magnitude.

Point-dot maps; their construction and use. Charts 42 to 45 illustrate the correct use of point-dot maps. In each instance an aggregate for the United States is distributed among the States — numbers of automobiles, numbers of dairy cows, dollars' worth of dairy products, and bushels of corn. Each map gives an instantaneous impression of comparative densities — the localities where there are relatively few automobiles, the localities where they are found in greatest numbers; and similarly for the other maps. Though it is something of a

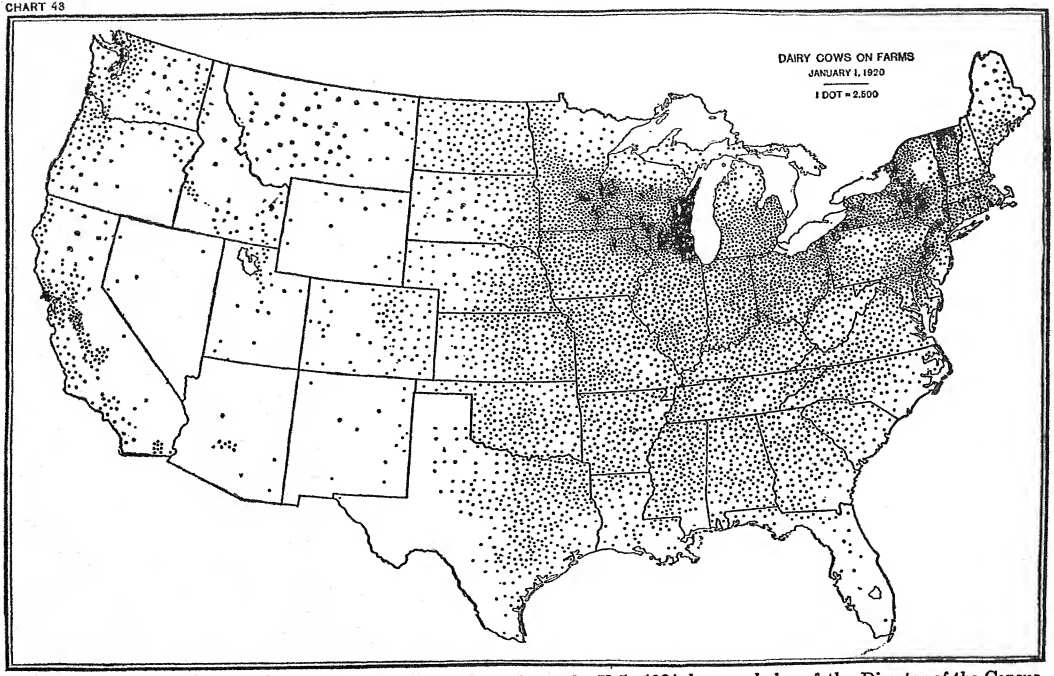
[1] See Day: *Statistical Analysis*, 219, quoting the United States Department of Agriculture.

CHART 42



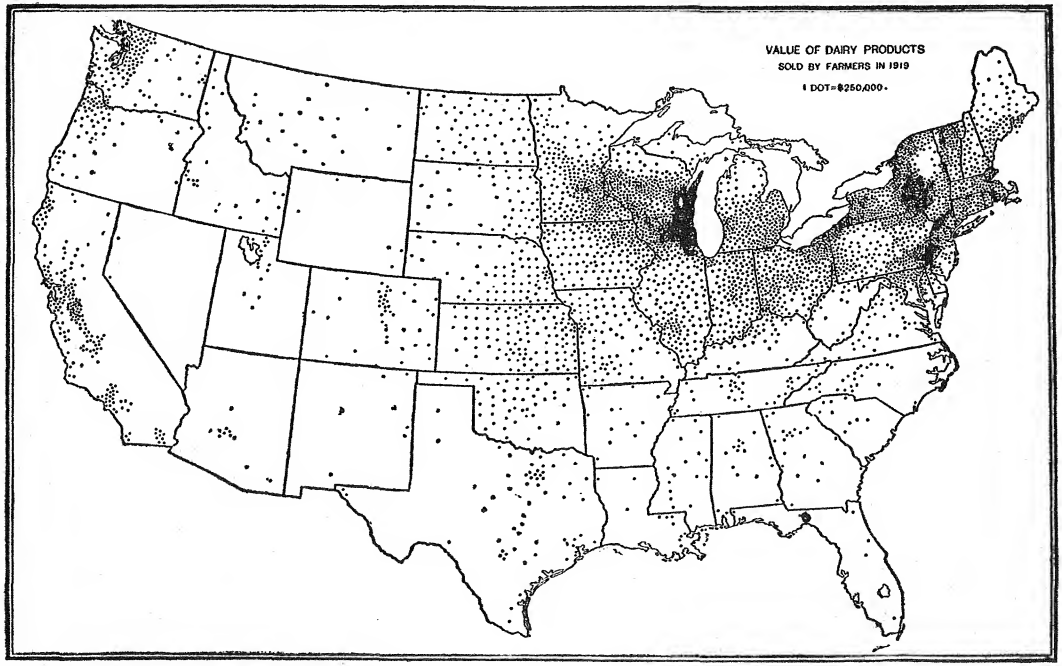
Reproduced from the *Statistical Atlas of the U.S.*, 1924, by permission of the Director of the Census.

CHART 48



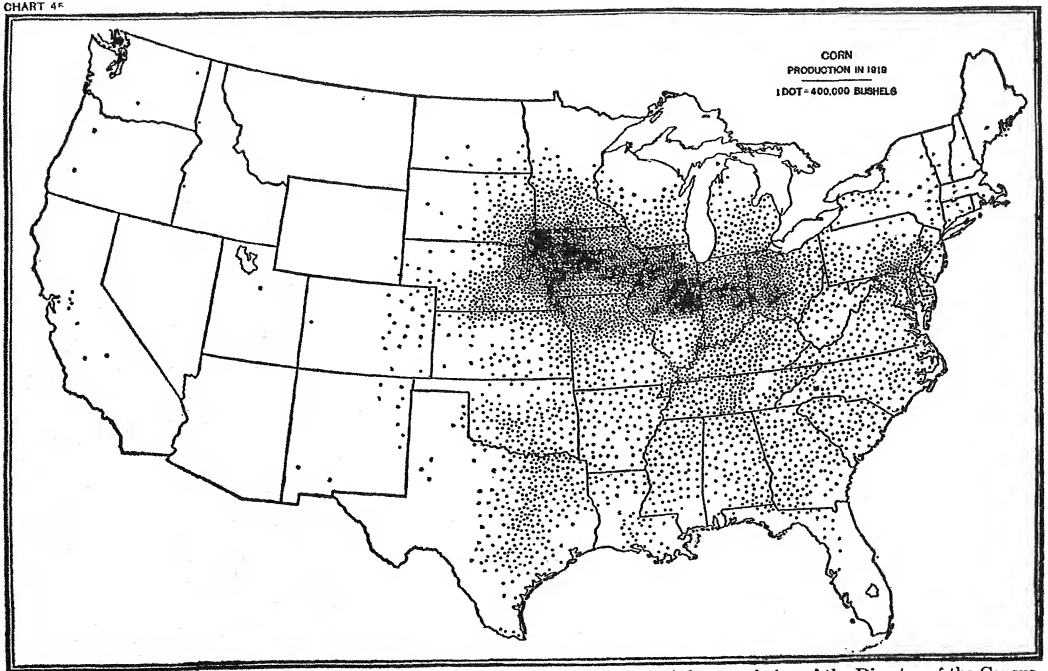
Reproduced from the *Statistical Atlas of the U.S.*, 1924, by permission of the Director of the Census.

CHART 44



Reproduced from the *Statistical Atlas of the U.S.*, 1924, by permission of the Director of the Census.

CHART 45



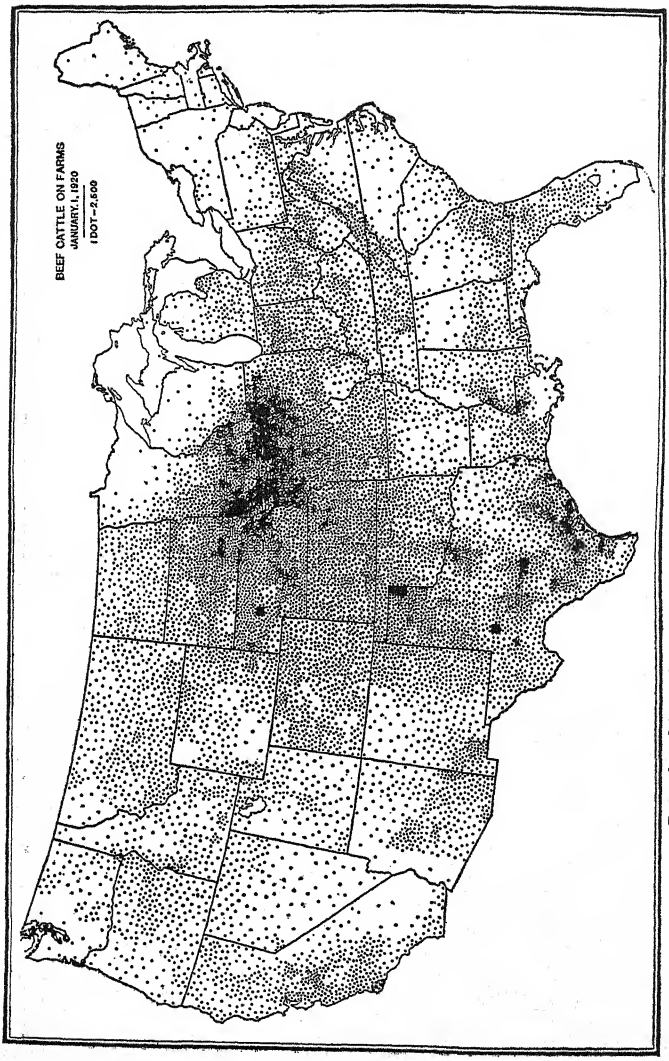
Reproduced from the *Statistical Atlas of the U.S.*, 1924, by permission of the Director of the Census.

routine task to construct this type of map because of the time required to insert the dots, there is nothing especially complicated in its make-up. The number of units of frequency represented by each dot must be determined with some care. It will be noticed in every successful point-dot map that there is at least one place in the map where the dots cluster so closely as to give the impression of almost solid black. The number of units of frequency to be represented by each dot must be determined by reference to these areas of greatest density and with the purpose in mind of producing the effect of nearly solid black therein. The dots should be uniform in size and small enough so that each one is essentially no more than a point on the map. Since the map is a representation of frequencies classified for a given set of area units, such as States in the United States, it is proper to assume that there is no knowledge of the distribution of frequencies within the boundaries of any given space unit, and the appropriate procedure, therefore, is to plot the dots uniformly over this area. That the dots are not uniformly plotted over each State in charts 42 to 45 is due to the fact that the original maps were constructed with the counties and not the States as units of area.

In the cases cited above, each map stands alone. The purpose has been to show density variations over the entire area of a map and to permit comparisons of these densities in its different parts. It is possible, however, to carry the uses of such a map beyond this point and to make comparisons of one map with another. The comparison may concern merely the relative distribution of the frequencies over the total area or it may involve in addition a comparison of total frequencies in the two maps. Suppose the facts dealt with the acreage of wheat and of corn in the United States and the data were available for each State. If these two maps were constructed in accordance with the principles outlined above, one dot might represent 500 acres of wheat in one map and 750 acres of corn in the other;

and by a comparison of the two it would be possible to tell at a glance how the acreage of each grain was distributed and to see whether the areas of greatest and of least density coincided in the two maps, and if they did not, to note significant differences. But it would not be possible from these two maps to say for which crop the greater acreage exists. The latter comparison is a difficult one at best, but if it can be done at all it must be done by making the units represented by the dots in the two maps equal or equivalent. In the illustration given, if the unit in each map is one dot = 500 acres, the relative distribution of frequencies can still be seen as before; but a considerably greater acreage of the one crop than of the other will also be made evident through greater frequency of the dots. Compare, for instance, charts 43 and 46. Dairy cows are most highly concentrated in Wisconsin, Minnesota, and a few Eastern States, principally New York; and the industry is mainly located in the eastern half of the United States. Beef cattle, however, are found in greatest numbers in Iowa and to the south into Texas, but are found in fairly large numbers all over the United States with the exception of the Atlantic Coast. Since also in each map one dot represents 2500 animals, the two maps can be compared as a whole or in individual localities in terms of absolute frequencies. Comparison of the two maps in their entirety indicates that there were more beef than dairy cattle in the country on January 1, 1920; and similar comparisons may be made for local areas: thus, more beef than dairy cattle in California; more dairy than beef cattle in New York and Pennsylvania, etc.

Comparisons of absolute frequencies are often difficult or impossible to make because of the different kinds of units involved or because of differing significance of the units. Thus one might for some purposes consider acres of wheat equivalent to acres of corn, but in economic significance they might differ greatly. Even bushels of wheat are not economically equiva-



Reproduced from the *Statistical Atlas of the U.S.*, 1924, by permission of the Director of the Census.

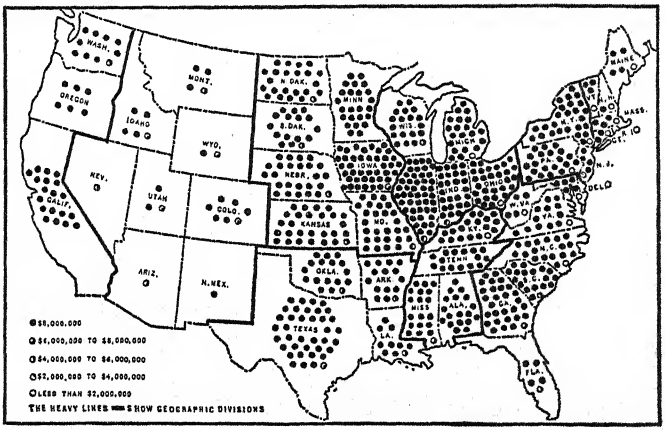
lent to bushels of corn, as their prices in the market testify. A still more difficult case arises in comparing beef produced and beef consumed, the former in terms of number of animals, the latter in hundred-pound units of dressed beef. A fair comparison can often be made in cases like the last one named if a reasonably stable ratio can be established between the two units. If one animal is equivalent on the average to 500 pounds of dressed beef and if one dot represents 2500 animals, then one dot in the dressed beef map should represent 1,250,000 pounds. By such methods as these comparability can frequently be obtained. The possibilities of the method do not extend, however, to such cases as comparing distribution of beef cattle with distribution of wheat production in terms of absolute frequencies; these two maps will indicate comparative density variations but not comparative absolute frequencies.

Large-dot maps; their construction and use. The use of the point-dot map is confined exclusively in this exposition to the presentation of data of spatial distributions of frequency, and its special fitness in this use is based on its capacity to portray density variations. Sometimes, however, interest in spatial distributions is centered not alone upon density variations, but also upon the matter of total frequencies. The purpose is to show the quantities of a given phenomenon in various localities, in such a way as to convey the impression of total or aggregate. Certain it is that the point-dot map, which shows relative densities so well, reports no results that can be put in terms of *how much*. The large-dot map, however, does this latter thing very well. It in turn is a poor medium for reporting densities, but it furnishes an excellent method of portraying the facts graphically in such a way that a fair estimate of totals is possible. The statement that the large-dot map reports *countable* frequencies gives its chief characteristic and its main contrast with the point-dot map.

Chart 47 illustrates the large-dot map. Generally speaking,

it is possible at a glance to tell whether one State contains more or fewer dots than another, and if one desires to make more accurate comparisons of two States, it requires but a moment to count the dots and compare results. It was because of this characteristic that the large-dot map was said to report *countable* frequencies. The only point of technique involved in the construction of this map lies in deciding the size of the dot and

CHART 47 ALL FARM CROPS—VALUE BY STATES: 1909



Reproduced from the *Statistical Atlas of the U.S.*, 1924, by permission of the Director of the Census.

the number of units of frequency to be represented by each dot. The dots, of course, are all uniform in size. They should be large enough so that there will be a relatively small number in any given area subdivision of the map. This is necessary in order that they may be counted, if desired. They should also be arranged in some orderly way in each area to facilitate counting; it is frequently possible to arrange them in rectangular form. The size of the dots having been decided upon, their unit representation of frequency is determined in such a way that in the area where they are densest there will be adequate room for

the necessary number and no room to spare. In chart 47, one dot represents $8,000,000 of value, and with this unit the State of Illinois is densely covered with dots — so much so that the State name had to be omitted. The partly shaded dots of course represent fractional parts of the unit. The lengthy legend at the lower left-hand corner of the map could be greatly shortened. It is scarcely necessary to do more than to state that one dot equals $8,000,000 and possibly to add that partly shaded dots represent fractional parts of a unit.

The occasions that call for the use of the large-dot map include all cases of spatial distributions of frequency where emphasis is placed upon actual numbers or aggregates and where the purpose is to afford an approximate count of these totals. This idea is generally uppermost in showing graphic representations of the value of industrial outputs in different geographic sections — the value of agricultural, manufacturing or mining products.[1] During the World War the United States Geological Survey constructed world maps for the use of various departments in Washington showing the location and estimated quantities of various metals of importance for war purposes. These maps used the large-dot principle for showing quantities and were unusually effective presentations.

Comparisons of large-dot maps may be made one with another, so long as the units in the different maps are the same or equivalent. Thus a series of maps similar to chart 47 might show values of manufacturing output, of mining output, etc., a single dot in each map representing $8,000,000 of product.

Construction and uses of cross-hatched maps. Cross-hatched maps are used to show spatial distribution of magni-

[1] The *United States Statistical Atlas* for 1924 uses cross-hatched maps for showing these facts by groups, also for the distribution of the Indian population by States, and the cotton ginned in each producing State by counties. See Plates 185, 252(2), 350(1), 352(3), and 364–369. The considerations developed above lead to the conclusion that cross-hatching is not as satisfactory a representation in these cases as is one or the other of the two types of dot maps.

tude variations, when it is appropriate to group the magnitudes in a small number of classes, usually not over six or eight. The occasions that call for such grouping usually involve data of rates or ratios given for each subdivision of area. Typical instances are population density (population per square mile) of States or counties, percentage increases in population, proportions of population falling into certain classes such as foreign-born, negro, gainfully employed, etc.; percentage of land in farms, percentage of improved land devoted to certain crops, sales per-capita population, per-capita mileage of paved road, etc. In showing the distribution of these data in geographic space, a clearer picture of the situation can ordinarily be obtained by showing a few well-defined magnitude classes into which the data fall than by attempting to show the actual magnitude for each spatial category. The latter is in danger of confusing the picture by too much detail and therefore of defeating its purpose. It should be emphasized that the attempt to display magnitudes graphically in space is a quite different matter from the simple comparison of magnitudes where the space factor is omitted. The latter is excellently shown by a series of bars starting from a common origin and arranged in the order of size; but as soon as it becomes necessary to show these magnitudes arrayed in space the possibility of a common origin for use in making size comparisons is lost; the possibility of detailed comparisons of the actual magnitudes is thereby lost and nothing remains but to show the spatial distribution of these size classes. But it should not be thought that this limitation of the graphic representation of spatial facts means that for want of a best method of representing the complete data a half-way or inadequate method is advocated. In the same way that a frequency distribution of magnitude variations[1] often gives a clearer picture of the character of the variation (or possibly of the *law of variation*, if such exists) than any possible

[1] See pages 90 to 94.

array of the individual magnitudes, so the magnitude groups of the rates and ratios here in question may convey a clearer conception of the facts than would result if the attempt were made to show each individual magnitude. Grouping magnitude variations is therefore not a second-best method adopted for want of a better but an independent procedure which stands on its own feet and produces a result attainable in no other way.

When the data involve, not relative frequencies or rates and ratios, but variations in magnitude of a measurable characteristic in space, such as the price of wheat, or rainfall, or temperature, a question may arise whether it is desirable to group the magnitude variations into classes for the purpose of showing them on a cross-hatched map. If they are not grouped, there is no satisfactory method of representing them on a statistical map, as has been shown;[1] but if grouping is permissible, the cross-hatched map may be used for them in the same way as for rates and ratios. As a matter of fact, actual usage of the cross-hatched map is confined largely to the latter. The *United States Statistical Atlas*, 1924, in over four hundred pages has not given a single instance showing grouped variations of a measurable characteristic in this way.

Charts 48 to 54 illustrate the methods of constructing cross-hatched maps,[2] the uses that may be made of them, and some of the pitfalls that need to be avoided if the fullest measure of success is to be attained. The first problem in technique involves the formation of magnitude classes and the question whether they shall or shall not represent equal intervals on the magnitude scale.

The demand by some authorities[3] that the range of the vari-

[1] See page 167.

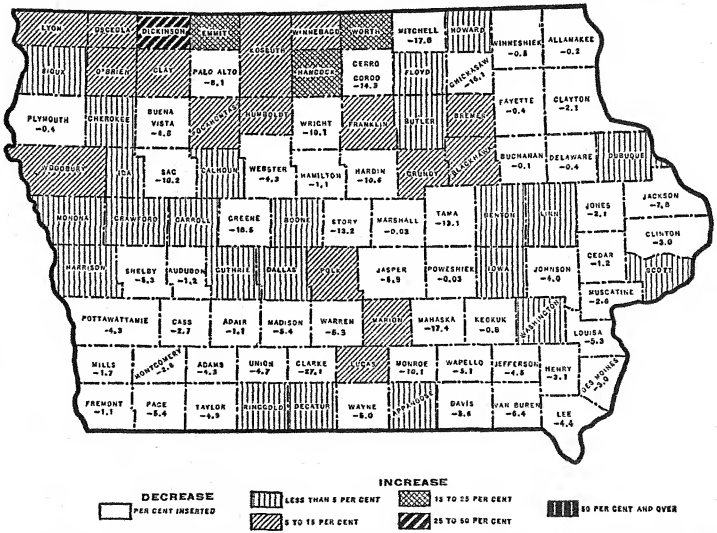
[2] The same distinctions as to magnitude that are made by density of cross-hatching can be made by variations in shading in color maps; but shading is a much more difficult process than cross-hatching and requires a more expensive equipment and a more skilled personnel to produce it. The Census Bureau has used color shading in its maps in the past, but in its statistical atlas published in 1924 gave it up entirely in favor of cross-hatching.

[3] See Day: *Statistical Analysis*, 217.

CHART 48

PER CENT OF INCREASE OR DECREASE IN POPULATION OF IOWA, BY COUNTIES: 1910–1920
RURAL POPULATION

Rural population is defined as that residing outside of incorporated places having 2500 inhabitants or more

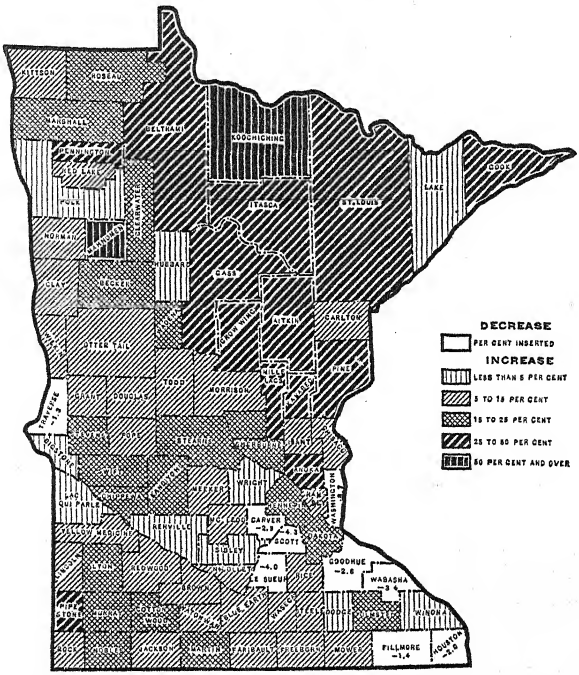


Reproduced from the *Statistical Atlas of the United States*, 1924, by permission of the Director of the Census.

Notice unequal intervals of magnitude classes.

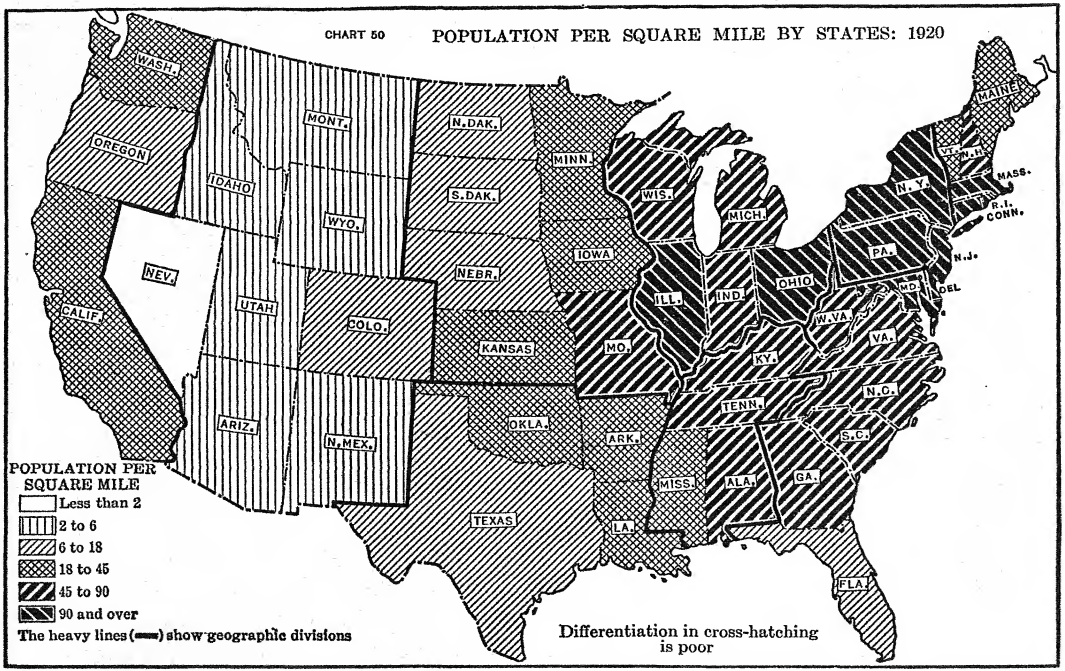
CHART 49

PER CENT OF INCREASE OR DECREASE IN POPULATION OF MINNESOTA, BY COUNTIES: 1910–1920

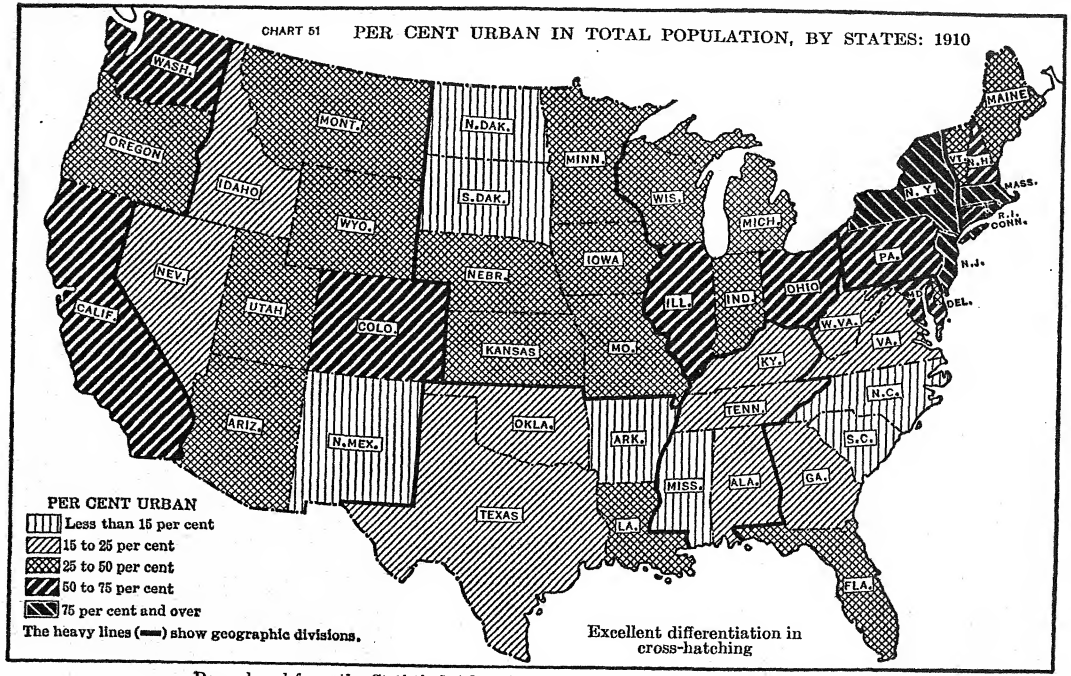


Reproduced from the *Statistical Atlas of the United States*, 1924, by permission of the Director of the Census.

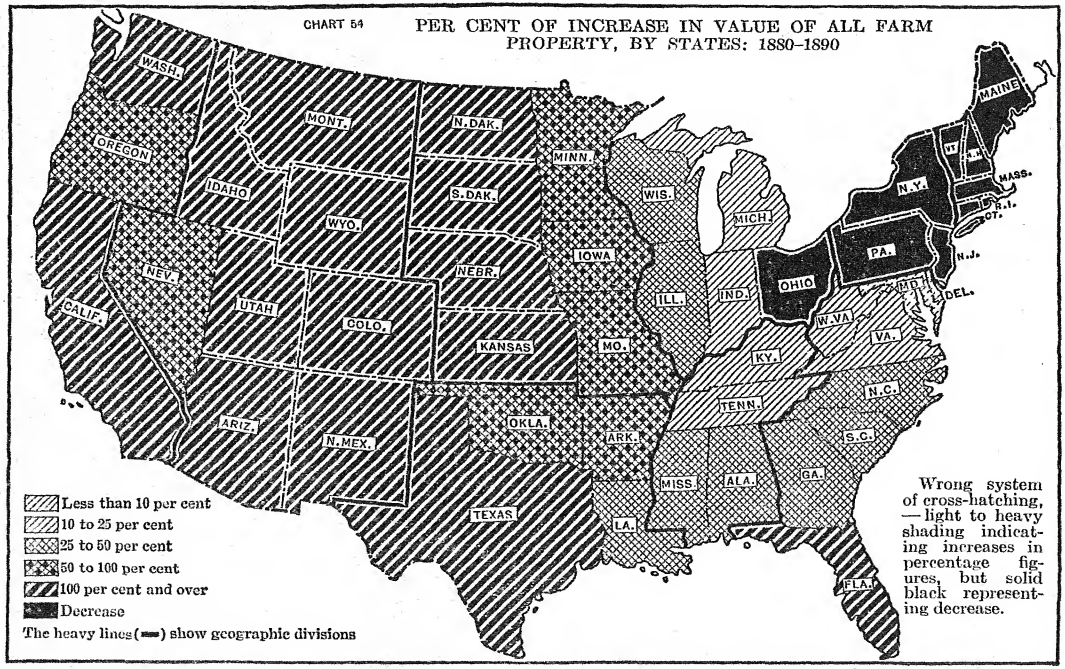
Notice unequal intervals of magnitude classes.



Reproduced from the *Statistical Atlas of the United States*, 1924, by permission of the Director of the Census.



Reproduced from the *Statistical Atlas of the United States*, 1924, by permission of the Director of the Census.



Reproduced from the *Statistical Atlas of the United States*, 1924, by permission of the Director of the Census.

able shall be divided into a number of equal intervals of magnitude for the purpose of forming the magnitude classes is based upon the thought that the situation here is identical with that in the construction of a frequency distribution of magnitudes. In the latter the search for uniformities in variation, for a law of variation of frequency,[1] necessitates magnitude classes of equal size throughout the range of the variable. But it is doubtful whether this conception is appropriate to the problem involved in distributing magnitude classes in space. In the latter case rather than contemplating the original rates, ratios, or magnitudes as reflecting any law of variation, instead of thinking of frequency as a function of size, the character of the data is probably more properly reflected by thinking of these magnitudes as being grouped around certain positions within the total range of variation; and these positions of concentration are determined to a large extent by the *heterogeneous* character of the observations *due to the space factor*. Take for instance the figures for per capita income tax, 1925, for the forty-eight States, Hawaii and the District of Columbia,[2] arranged here in the order of magnitude:

$0.24	$1.13	$1.32	$2.69	$8.57
.25	1.14	1.74	2.73	8.61
.28	1.20	1.87	3.01	8.88
.33	1.23	2.03	3.23	9.00
.46	1.26	2.07	3.45	9.14
.55	1.27	2.31	4.09	9.90
.76	1.28	2.35	4.69	11.12
.86	1.29	2.39	7.45	11.68
.93	1.30	2.42	7.74	22.59
1.12	1.31	2.60	7.76	22.83

Cases are found at reasonably frequent intervals up to the figure $4.69; but a large interval separates this from the next

[1] See pages 91–92.

[2] Data from *United States Statistical Abstract*, 1928, page 187.

figure, $7.45. There are no cases in the $10 interval and none between $11.68 and $22.59. If these figures are thrown into a frequency distribution with a one-unit interval, as shown herewith, the result is a very irregular distribution, not at all similar in character to the distribution of weights of 1000 Freshmen shown on page 91. The income tax figures do not produce a distribution of the typical sort with frequencies falling off gradually on either side of a position of maximum frequency and this is due to the fundamental nature of the income tax data. The grouping of these figures for purposes of showing their characteristic features in a statistical map must take this peculiar quality of the figures into account. The grouping is determined by the data and should not be controlled by the rigid requirements of a frequency distribution where the characteristics of the latter are not present in the data. A classification of the above figures into the following groups better suits the character of the data than does the frequency distribution in one-unit classes:

Per capita Income Tax	Number of Cases
0.0–	9
1.0–	14
2.0–	9
3.0–	3
4.0–	2
5.0–	0
6.0–	0
7.0–	3
8.0–	3
9.0–	3
10.0–	0
11.0–	2
12.0–	0
13.0–	0
14.0–	0
15.0–	0
16.0–	0
17.0–	0
18.0–	0
19.0–	0
20.0–	0
21.0–	0
22.0–	2
	50

Per-capita income tax
Under $1.00
$1.00 and under $2.00
2.00 and under 5.00
5.00 and under 12.00
12.00 and over

The conception of the frequency distribution is appropriate to the variation of magnitudes within a homogeneous group of individuals; but death rates, or density of population or proportions of foreign-born, etc. — data given for different space categories — are necessarily heterogeneous in character; they are not distributed in such a way as to reflect any law of variation, and it is therefore placing a wrong emphasis upon such data to distribute them in groups or classes appropriate to homogeneous observations. A result more in conformity with the character of the data is attained by grouping as above in classes that indicate real points of concentration of the observations, and this endeavor will not ordinarily lead to equal-sized class intervals, even though it may do so in particular instances.

It is a good plan, therefore, in deciding upon the interval to be represented by each class or group to array in the order of size the entire list of magnitudes involved and then to select the appropriate groupings from this array. Note that in chart 48, rural population, all counties showing population decreases between 1910 and 1920 are included in one group, although the decreases vary from .03 per cent to 27.1 per cent. The purposes of this map are better served by showing all the decreases in one group and increases in five groups than would have resulted from an attempt to divide the range of variation so that six classes of equal size would have resulted.

When the magnitude classes have been decided upon, it is necessary to select the type of cross-hatching that is to represent each class. There is one rule here that admits of no exception — the cross-hatchings must indicate variations in magnitude. The heaviest shading or the most nearly black shall represent the largest magnitude class, the lightest shading the smallest class, and proper gradations for the intermediate classes. Herein comes a difficult technical problem, the problem of making these distinctions really stand out in the map. To be sure, it is a problem in draughtsmanship, but the drawing

can easily fail of its purpose if the statistician's point of view is not kept uppermost. It is not a question merely of differing kinds of cross-hatching, but rather of clear-cut distinctions of magnitude. Compare, for instance, charts 50 and 51. In chart 50 the distinctions in cross-hatching are poor; it is somewhat difficult to say whether the *45 to 90* class shows less black than the class *90 and over*; and it is safe to say that no differentiation in magnitude has resulted as between the class *2 to 6*, and the class *6 to 18*. In chart 51, however, the cross-hatching is excellently done, from the statistician's point of view; the smallest class is the lightest and the gradations from this to the largest perfectly reflect the size-classes of the variable. Chart 54 shows a map that violates the rule stated above; decrease in value is shown by solid black and the map is subject to serious or possible misinterpretation of the facts for this reason.

Comparisons can sometimes be made of one cross-hatched map with another. Corresponding facts may be shown for different census periods, such for instance as the proportion of foreign-born by States, 1910 and 1920; or the per-capita income or wealth by States for the same dates. For such comparisons it is necessary that the same grouping of the variable be used for each map and the same cross-hatchings represent corresponding groups in each map. Charts 52 and 53 violate this requirement, for while the groupings of the variable are the same in the two maps, the systems of cross-hatching differ in three out of seven of the classes. Comparison of the heavy diagonal black lines on the two maps for instance involves comparison of the *62½ to 75* class in one case with the *75 to 87½* class in the other. Each map individually is correctly drawn but they cannot be used for comparison one with the other.

Note: Other uses of shaded or cross-hatched maps are frequently found, but they are of a non-statistical nature and therefore deserve little notice here. One case shows qualitative differences, generally two or three, sometimes more, for different subdivisions of area.

Thus a workmen's compensation map of the United States may show in black the States that have no compensation laws, in light cross-hatching States that have laws but no State fund, and in clear white those States that have laws and a State fund. Another very valuable type of cross-hatched or shaded map, non-statistical in character, may be called the location map. Illustrations are maps with black shading showing the location of coal fields, petroleum fields, the location of various metal deposits, etc.

EXERCISES

I

Density comparison, number of concerns in business and number of failures. Data in *United States Statistical Abstract*, 1928, page 313.

Construct a point-dot map showing number of concerns in business in each State, 1926, and another map showing number of failures, 1926.

(a) For comparing absolute densities make one dot represent the same number of units in each map.

(b) For comparing relative densities, make one dot represent the same proportion of the total in each map. There being approximately one hundred times as many concerns as failures in the United States in 1926, a *relative density* comparison may be made in the two maps by making one dot represent one hundred times as many units in the business concerns map as in the failures map.

II

Comparison of aggregates (countable frequencies) in two maps. Data of I above.

Construct two maps from the above data using large dots to show frequencies. For comparison of absolute numbers in the two maps one dot should represent the same number of units in each.

III

Per cent of firms failing by States, 1926 and 1927. Data in *United States Statistical Abstract*, 1928, page 313.

(1) Construct a cross-hatched map to show the above facts for 1926.

(2) Construct a cross-hatched map to show the above facts for 1927 in such a way that comparison can be made with the 1926 map.

(Note contrast of this method with that of I (2) above.)

IV

Critical examination of statistical maps. Data in *Statistical Atlas of the United States*, 1924.

Select any of the large number of statistical maps given in the *Statistical Atlas* and criticize them individually and, where comparisons are made or can be made, in comparison one with another.

INDEX